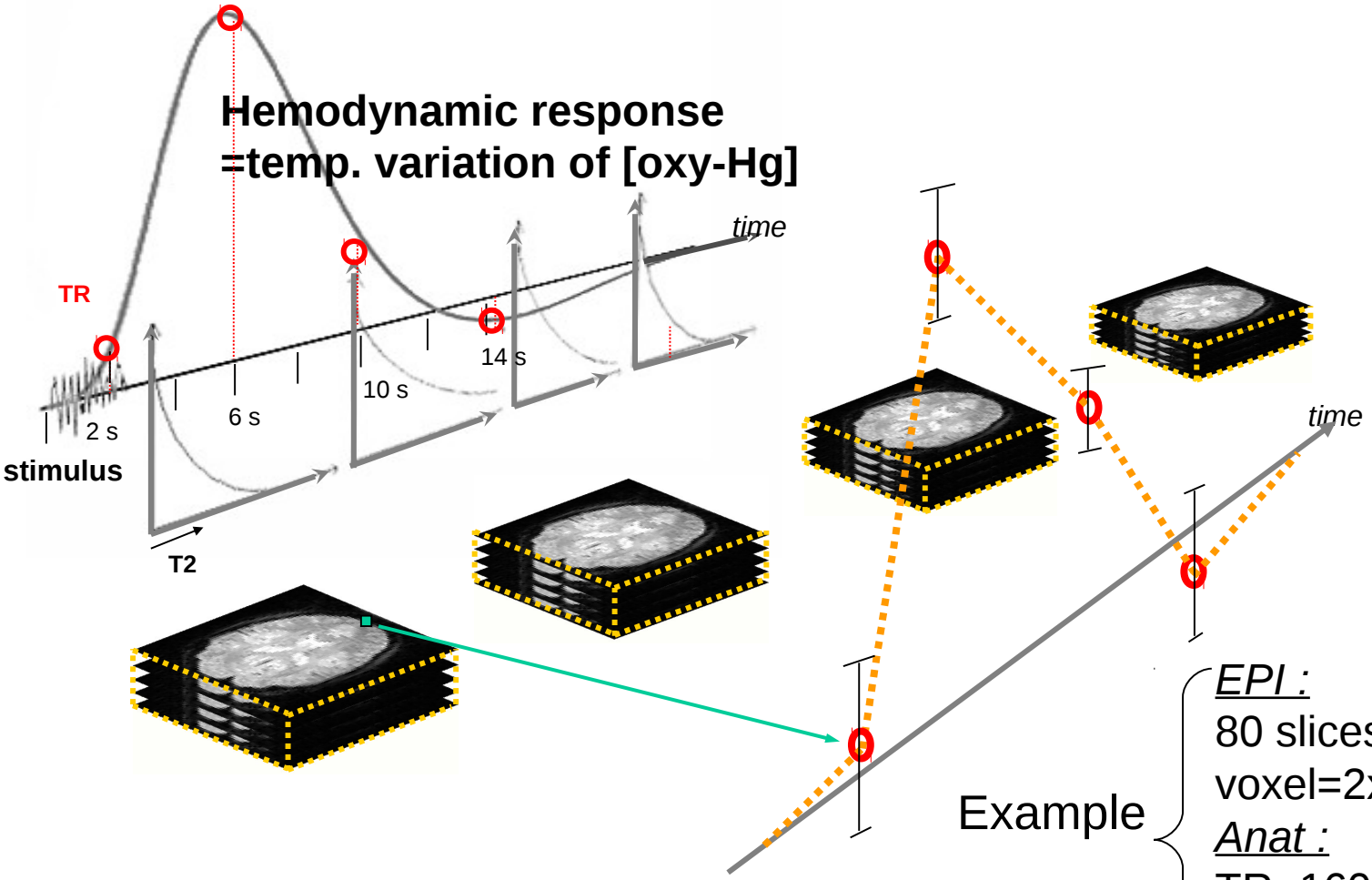
Part III Machine Learning for functional Neuroimaging

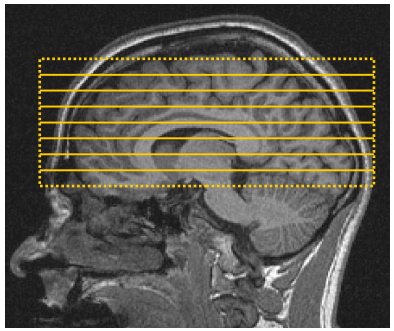
Outline

- The standard *multivariate pattern analysis* setting
- Identification of predictive features
- Mapping *and* predicting: Spatial Regularization
- Toward big data analysis

Introduction: functional Magnetic Resonance Imaging

The BOLD response



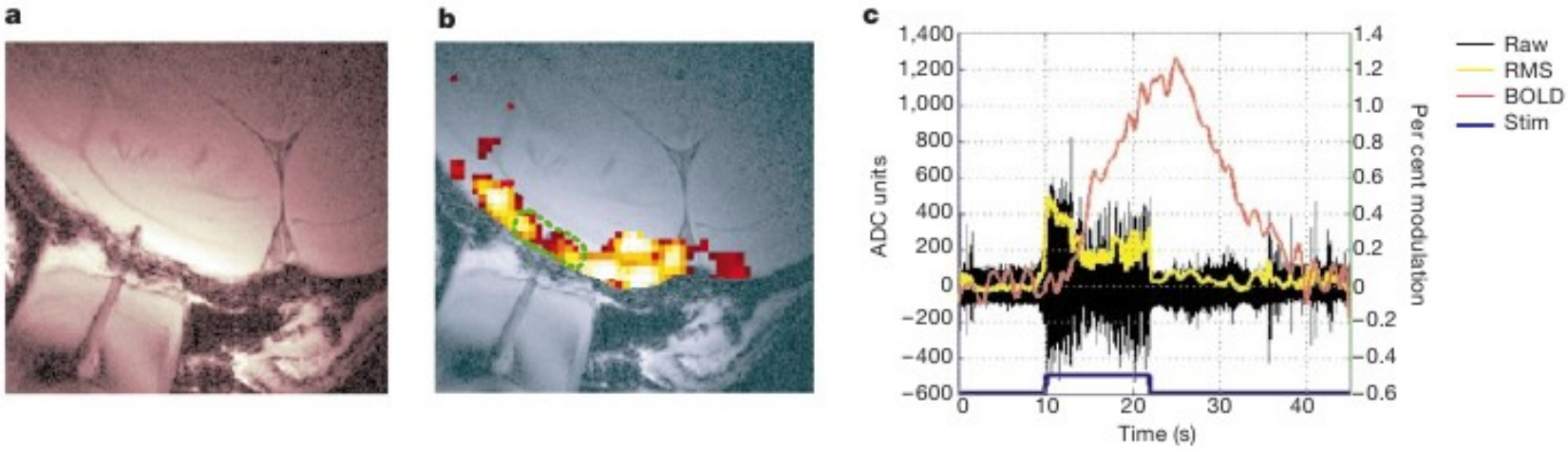


Example

EPI :
80 slices, TR=2000 ms, TE=30ms,
voxel=2x2x2mm, 96x96 matrix
Anat :
TR=1600ms, FOV=256x256mm2,
256X256 matrix, 192 slices
Slice thickness 1mm

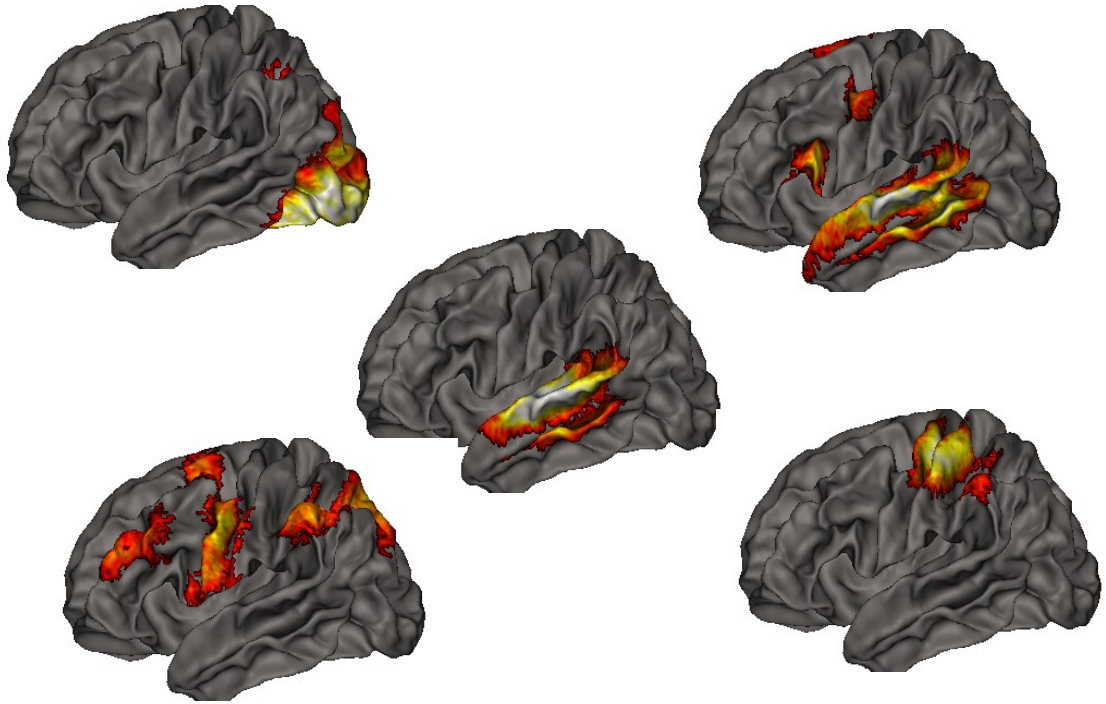
Human BOLD imaging

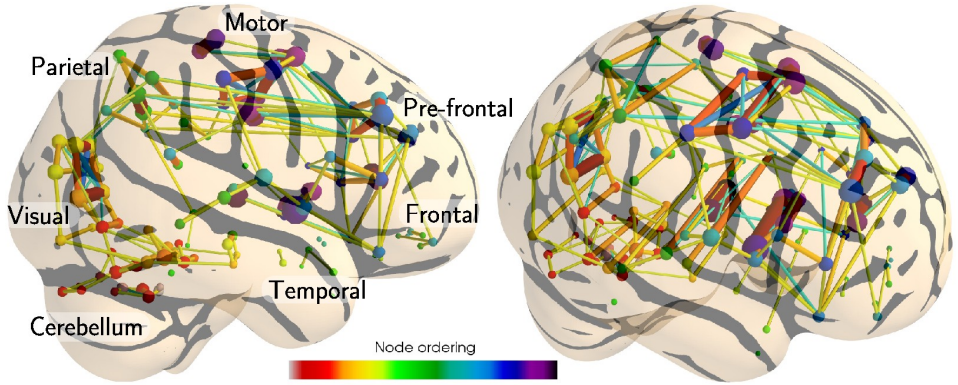
- BOLD response approximately linear in the stimulus function
  - Application of simple linear model for data analysis [Friston et al. 1995]
- BOLD signal highly correlated with LFPs [Logothetis et al. Nature 2001]
- High spatial accuracy (~2mm) [Ugurbil et al. NeuroImage 2007]
- Poor temporal resolution, no consensual model on the signal



Human BOLD imaging

- fMRI used intensively to map cognitive functions in the human brain.
- BOLD correlations across regions used to study functional connectivity
- Combining the signals across regions can be used to fit accurately the subject's behavior

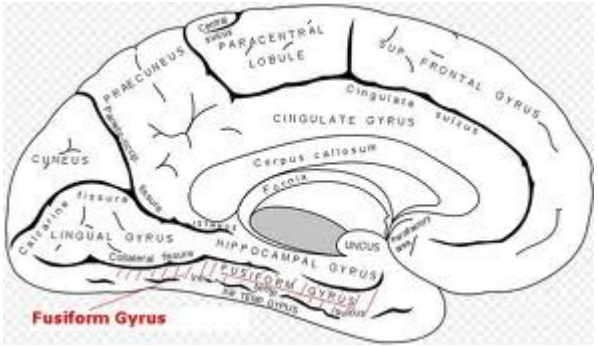




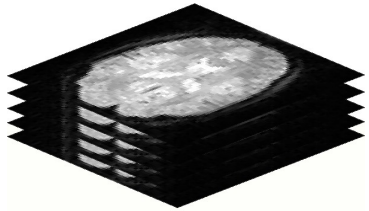
Machine learning for cognitive neuroimaging : decoding and encoding models

- Statistical analysis of fMRI data
- encoding models: mapping stimulus features to brain activations
- decoding models: predicting behavior from activation maps / “Vanilla MVPA”

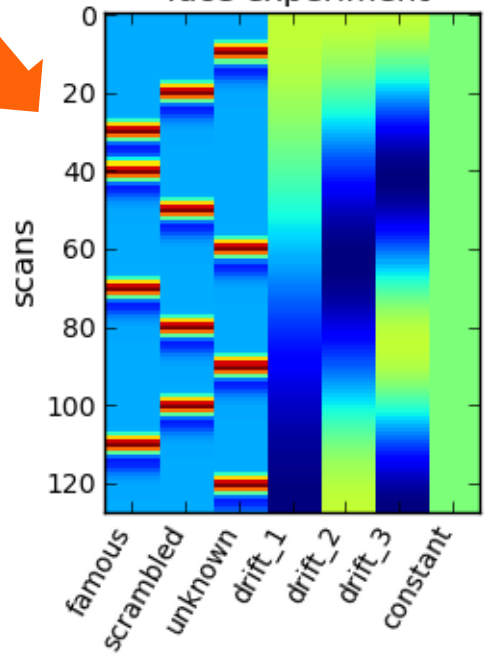
fMRI data from acquisition to analysis

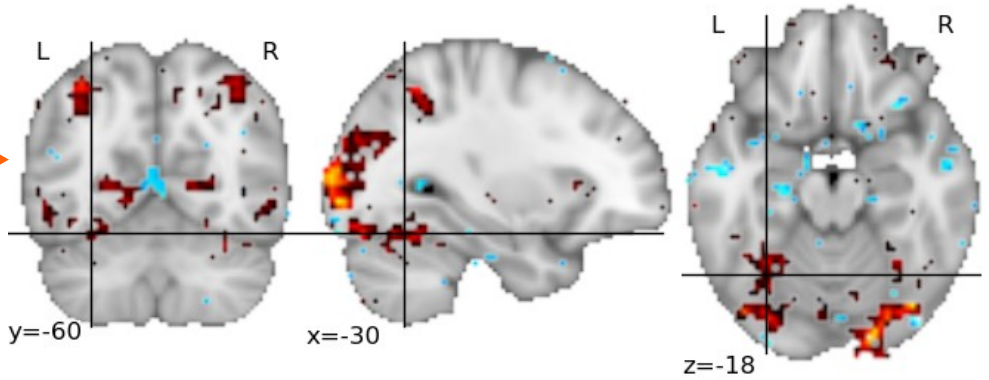


Complex metabolic pathway



Block design matrix face experiment





Statistical inference

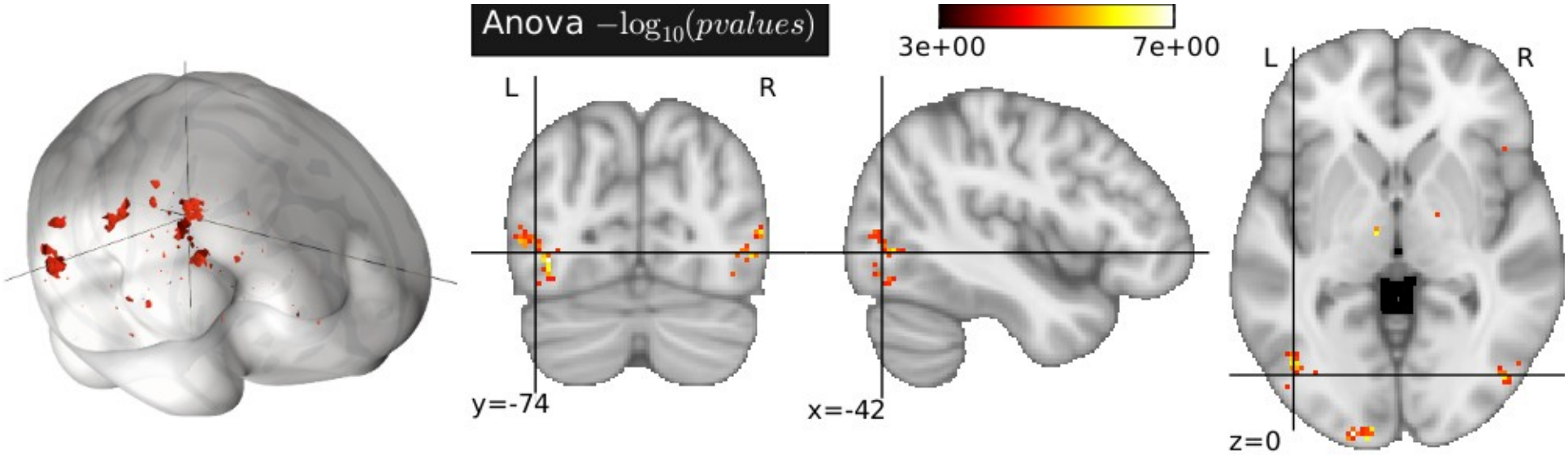
Question 1 : Is there any effect of the paradigm on the data ?

“omnibus test”

Can I discriminate between the two conditions ?

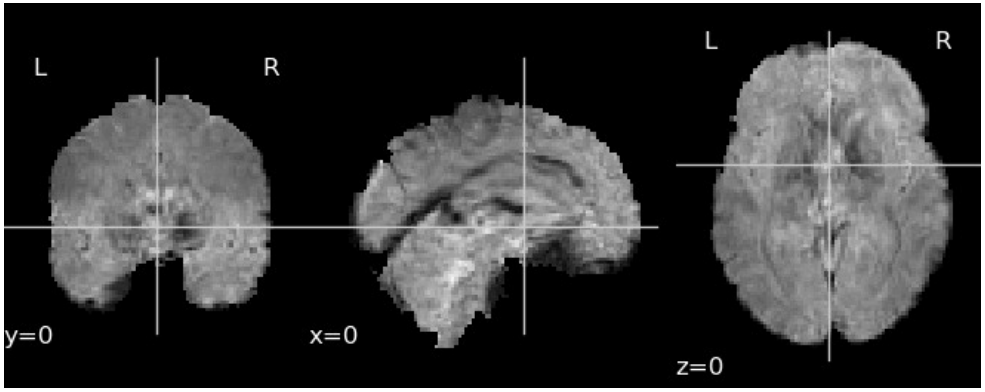
Question 2 : What regions actually display an effect of the paradigm ?

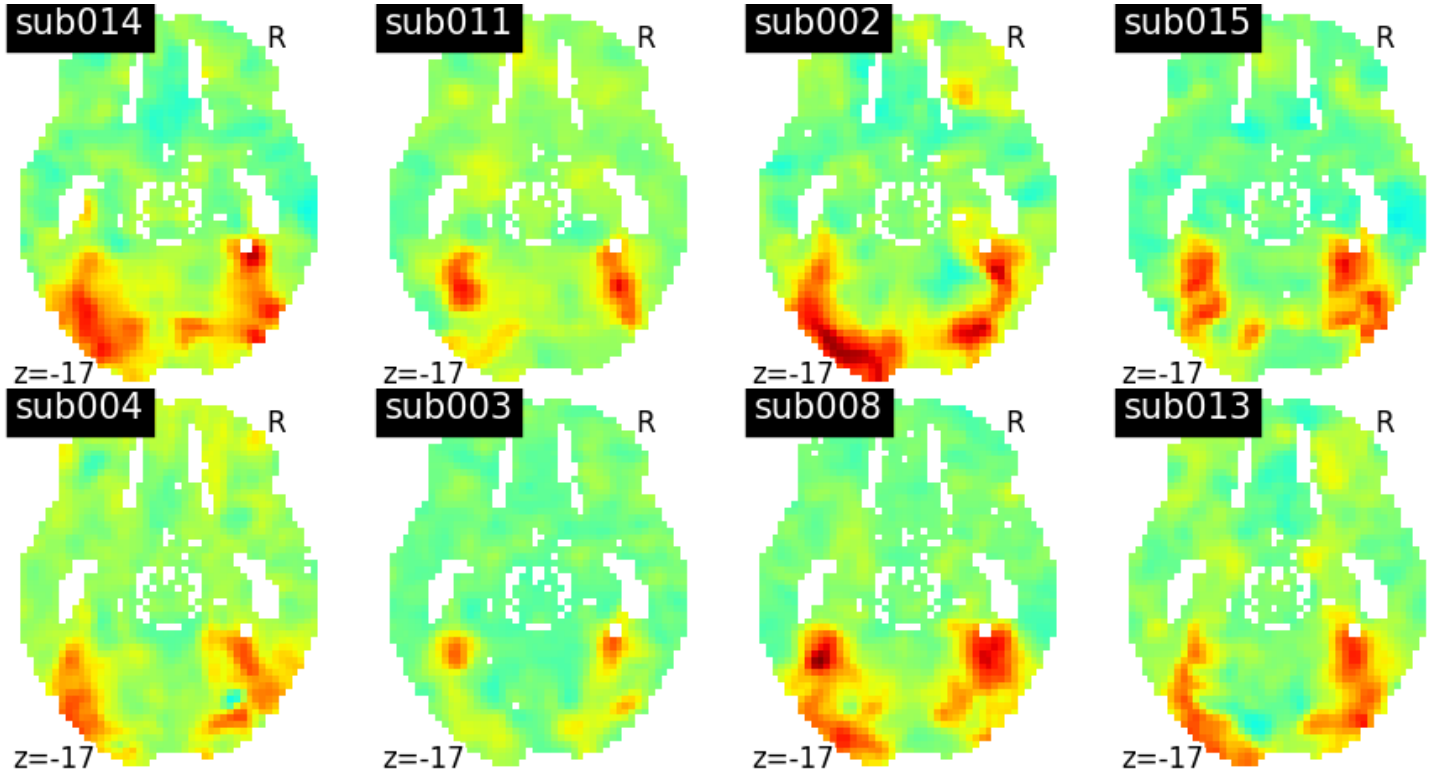
What are the relevant features ?



Common problems

- Low SNR
- High variability between subjects
- Small sample sizes





Common problems: sample size

Nature Reviews Neuroscience | AOP, published online 10 April 2013; doi:10.1038/nrn3475

nature REVIEWS

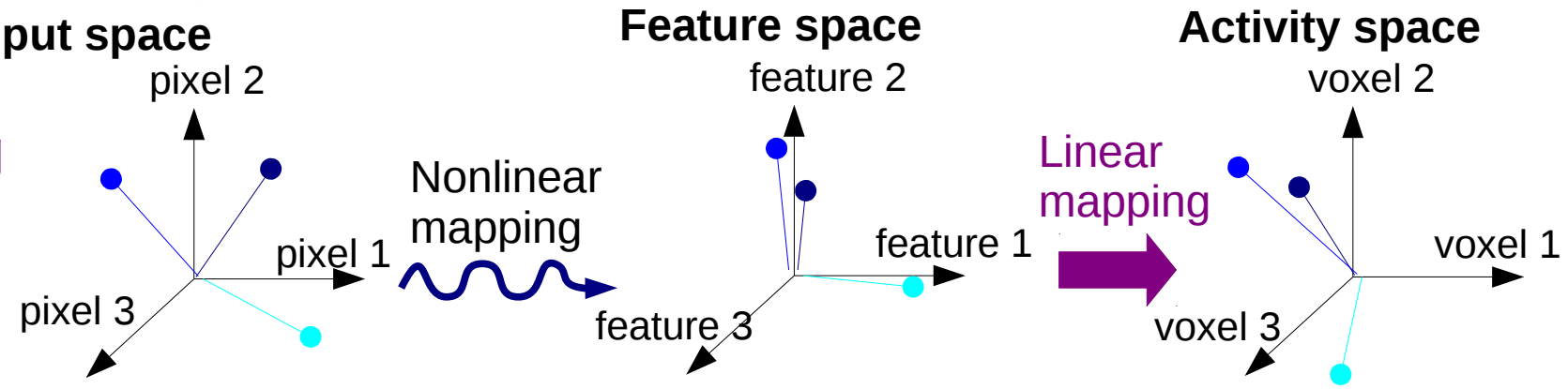
Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

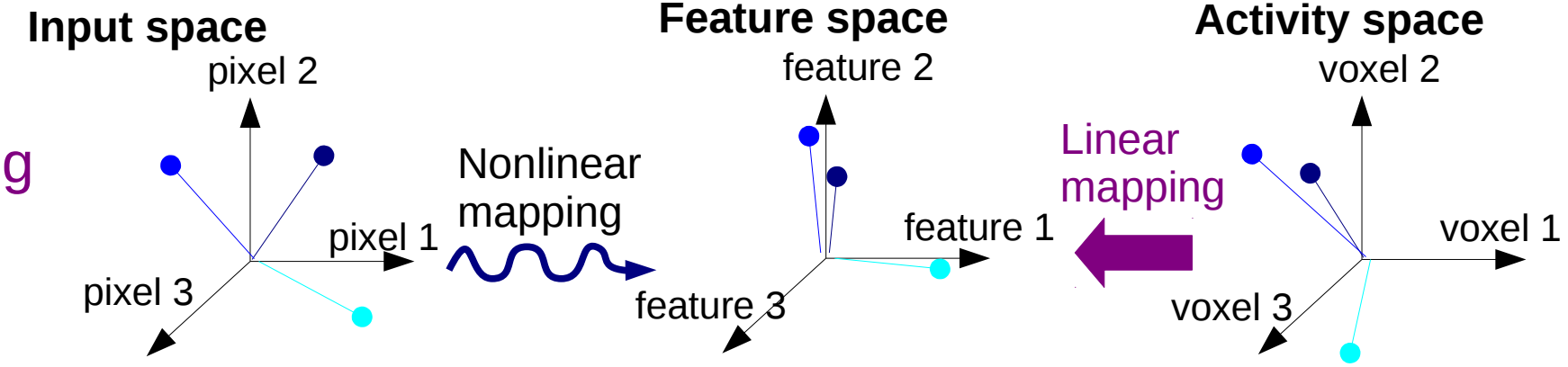
Small number of subjects → Lack of power
- → many false negatives
- → results are weakly informative
- → reported effects are inflated

All these issues are present in MVPA
Reflected in classifier instability, variable performance

encoding model



decoding model



[Naselaris et al. NeuroImage 2009]

Encoding visual stimuli

- Activation patterns encode some information about the stimuli
- Population receptive field
  - Location-specific (depends on where you are on the cortex)
  - Feature-related (position, contrast, speed, color...)

$$\mathbf{y} = \phi(\mathbf{X})\beta + \varepsilon$$

e.g. ϕ = wavelet transform

Signal in a voxel

Input stimulus

Voxel-specific response

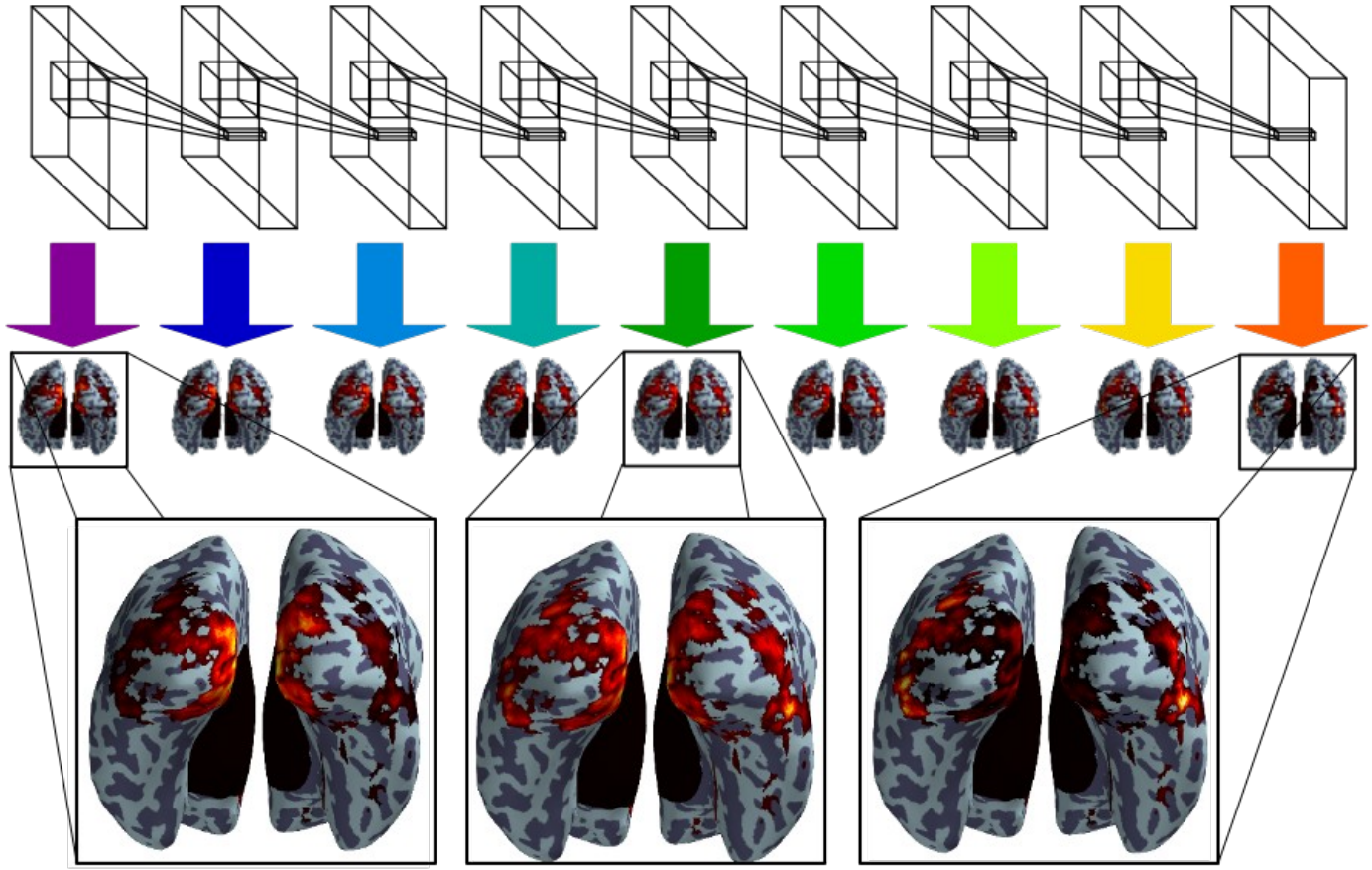
Encoding visual stimuli

Create Features

Convolution model: Feedfoward model of vision for object recognition

Comparison with fMRI

How well are brain regions explained by the features ?

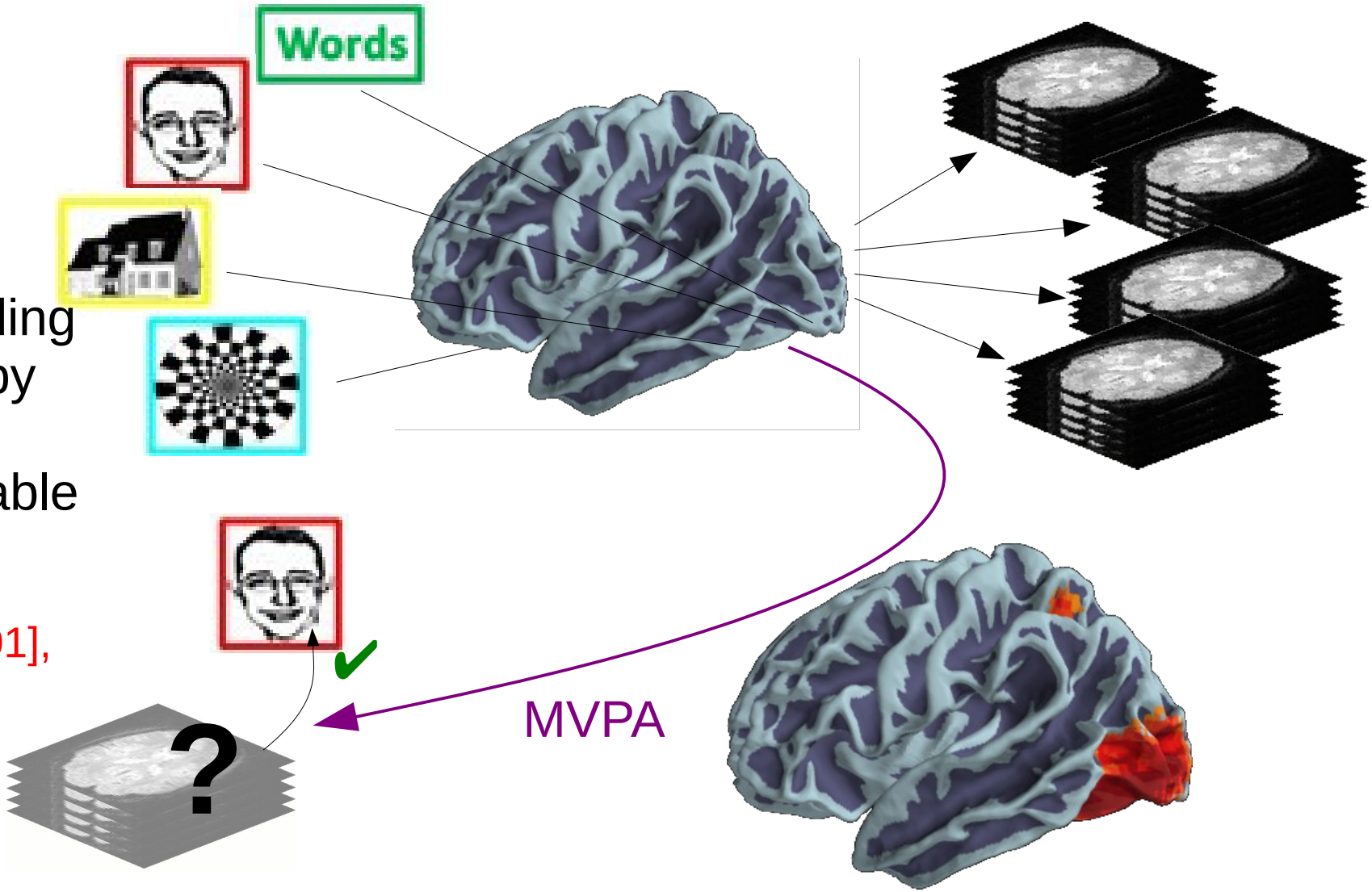


[Eickenberg et al. submitted]

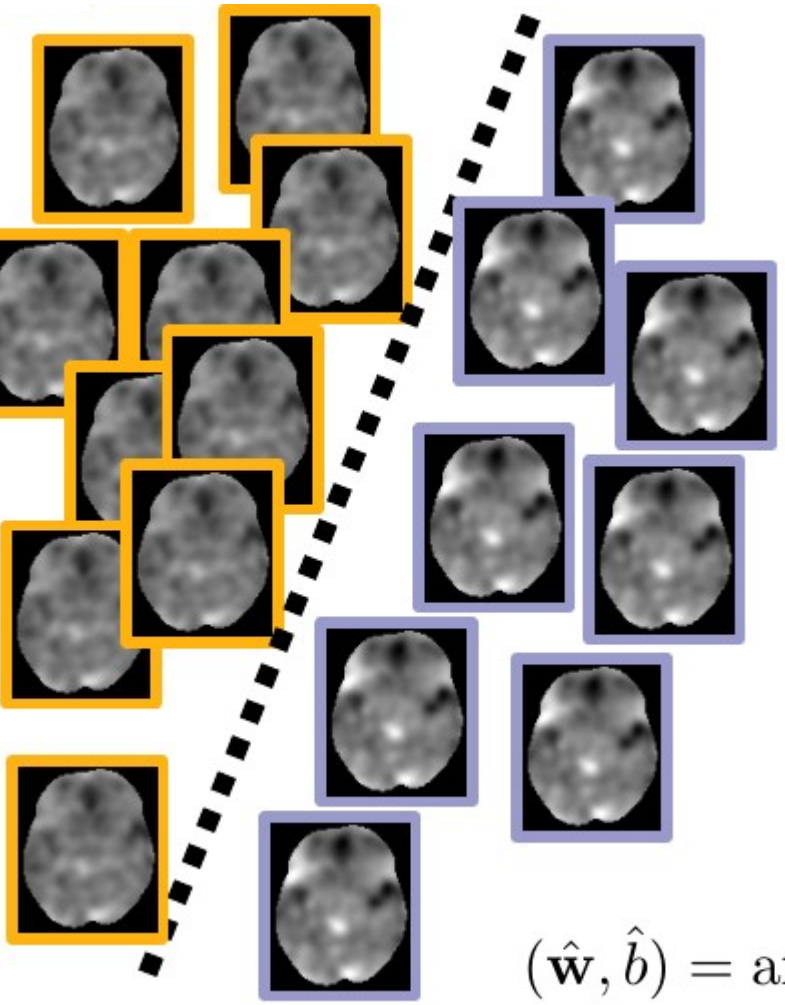
FMRI decoding: Reverse inference

Aims at decoding brain activity by predicting a cognitive variable

[Dehaene et al. 1998],
[Haxby et al. 2001],
[Cox et al. 2003]



Training classifiers for FMRI data



Main setting:
features = the voxel-level BOLD contrast
Target = conditions (discrete)

- Given x in R^p, (fMRI volume with p voxels), predict a label y in {-1, 1}
  i.e. ▮ or ▮

  or better the class probability
  Proba(x = 1|y)

$$(\hat{\mathbf{w}}, \hat{b}) = \operatorname{argmin}_{\mathbf{w},b} \sum_{i=1}^{n} \log\left(1 + \exp\left(-x_i(y_i\mathbf{w} + b)\right)\right) + \alpha\|\mathbf{w}\|_2^2$$

Evaluation of the decoding

Prediction accuracy

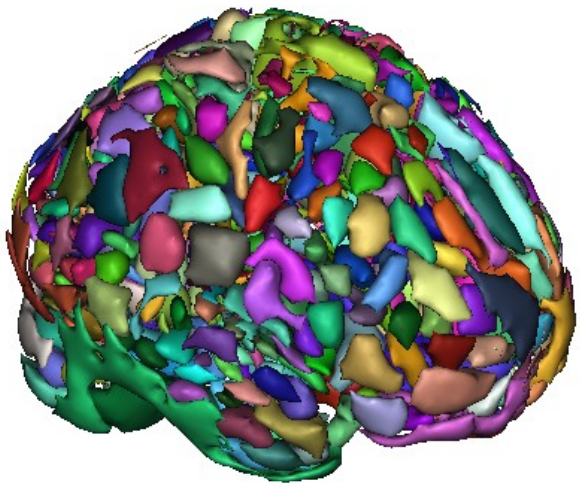
→ Quantify the **amount of information** shared by the pattern and y.

Layout of the resulting maps of weights:
The discriminative pattern characterizes brain processes / biomarkers

Common hypothesis = segregation into functionally specific territories

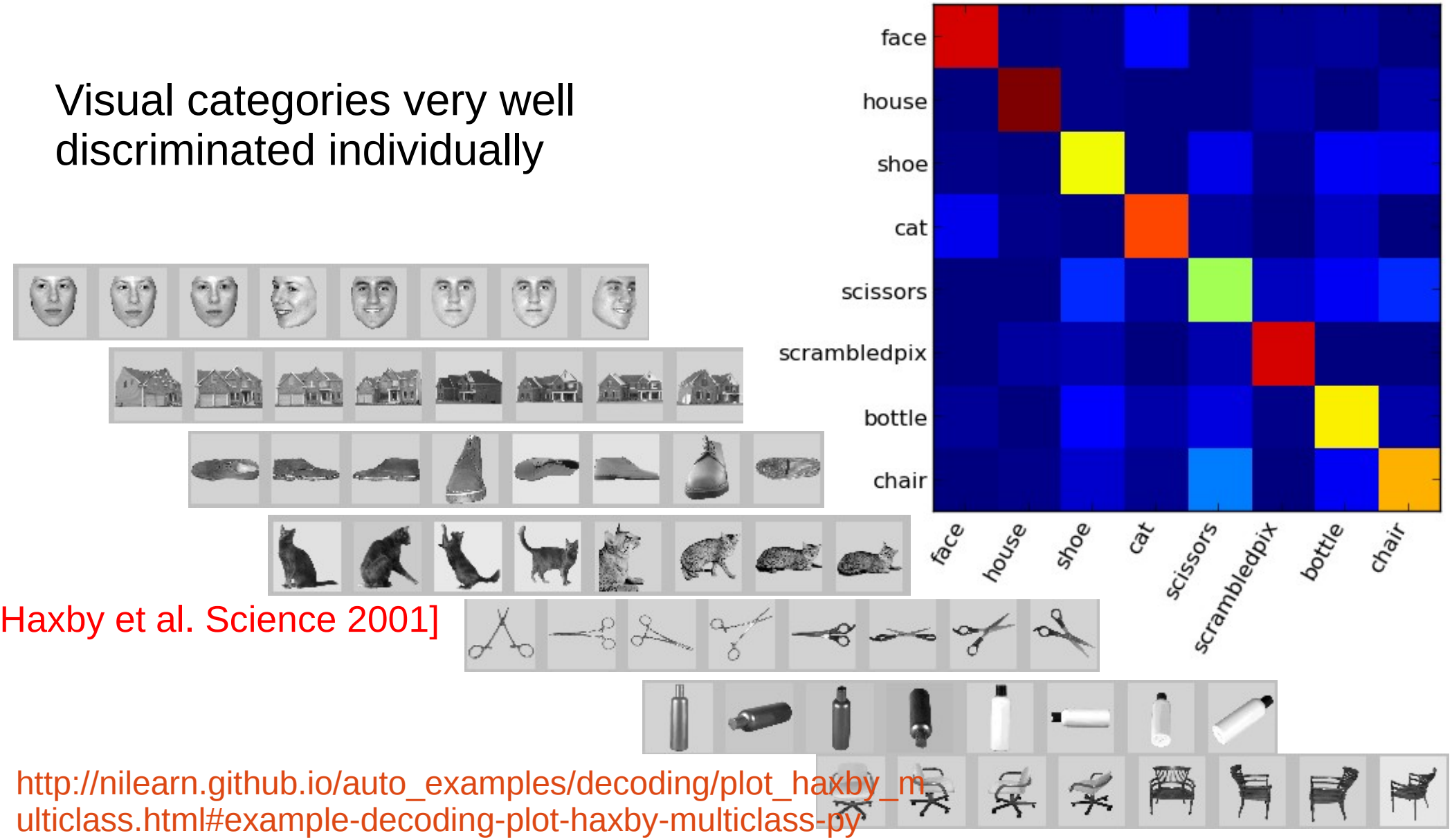
→ **sparse**: few relevant regions implied

→ **compact structure**: grouping into connected clusters.



Decoding visual categories

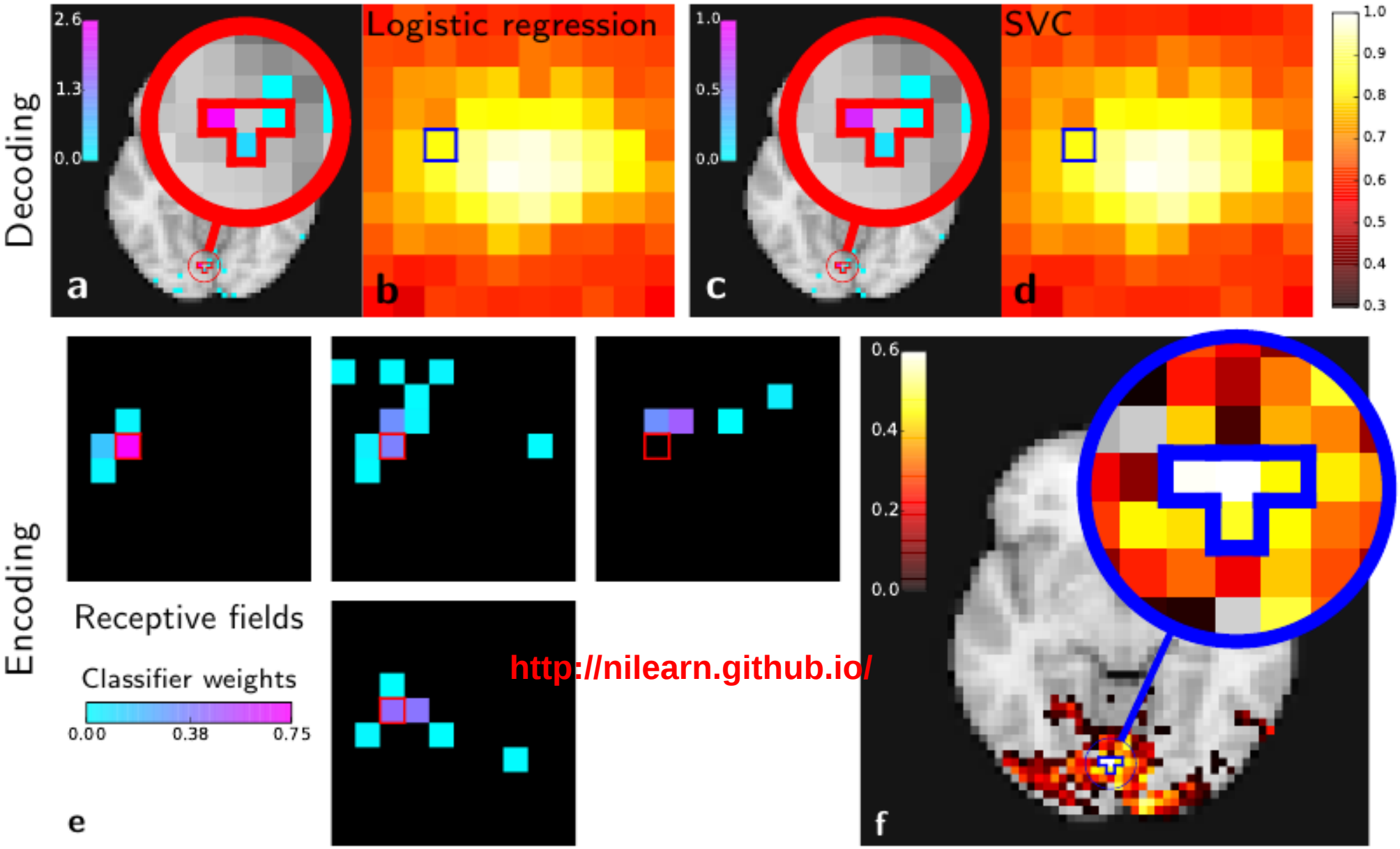
Visual categories very well discriminated individually



[Haxby et al. Science 2001]

http://nilearn.github.io/auto_examples/decoding/plot_haxby_multiclass.html#example-decoding-plot-haxby-multiclass-py

Encoding and decoding: do it yourself !



More topics

- Separate GLM approach improves classification accuracy [Mumford et al. NIMG 2012]
- Data-driven hrf estimation alos yields more accurate models [Pedregosa et al. Nimg 2014]

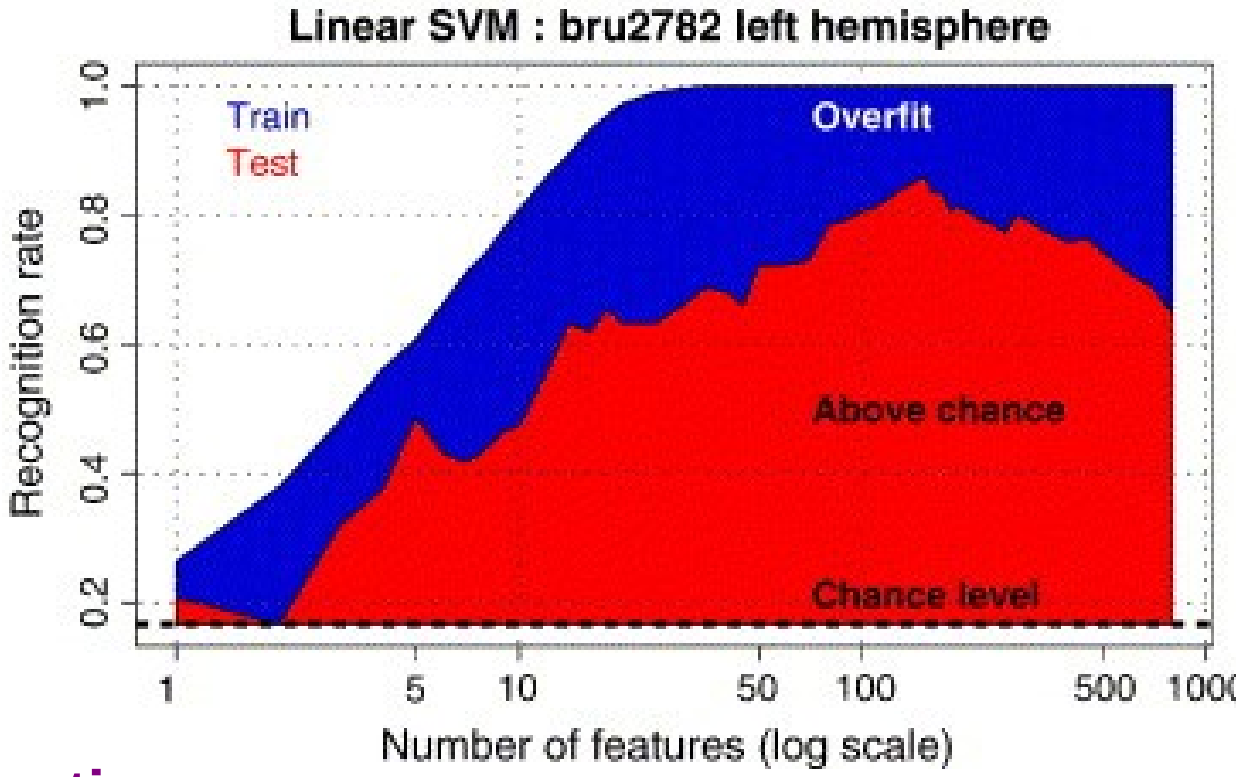
Region selection and Identification

- Spatial models in MVPA
- Identification : Recover the truly associated features
- Ill-posedness of the problem. Current solutions.

Curse of dimensionality in MVPA

Problem: p ≫ n

- Overfit the noise on the training data



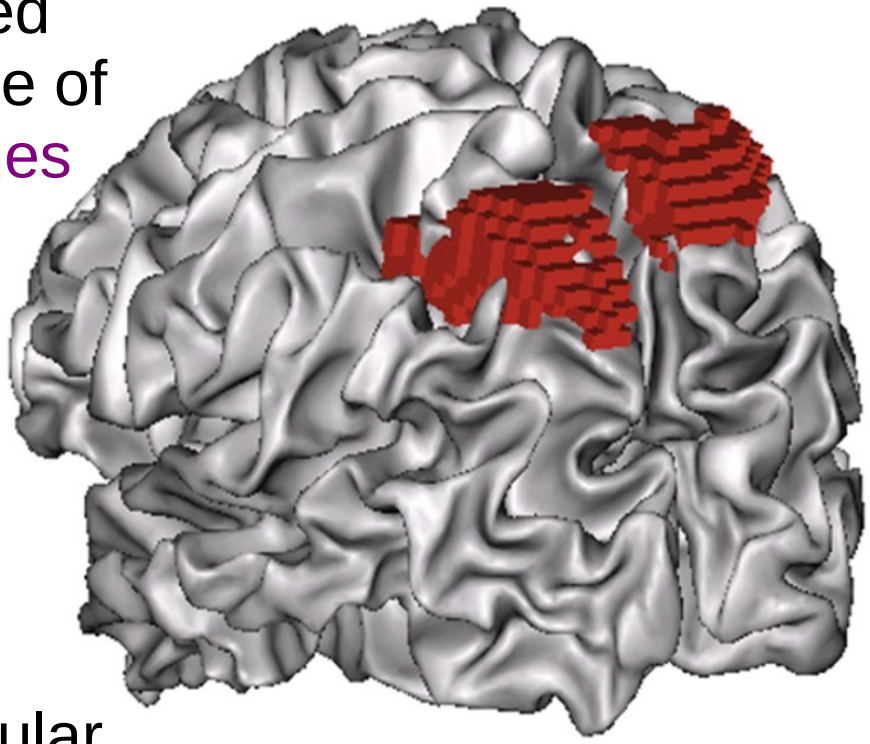
Solutions

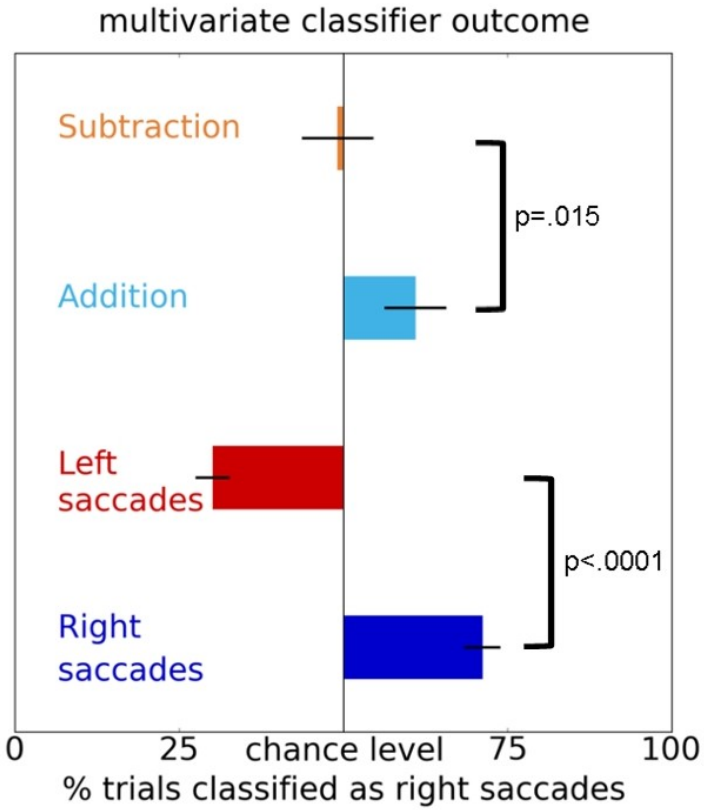
- **Prior region selection**
  - prior-bound result
- **Data-driven feature selection**
  - Univariate methods (Anova)
  - Multivariate methods
- **Regularization** (e.g. Lasso, Elastic net)
  - Shrink w according to your prior

ROI-based analysis

Example:

- Selection of results involved in performance of ocular saccades
- Succesful prediction of mental arithmetics
- Same neural populations involved in ocular saccades and arithmetics ?



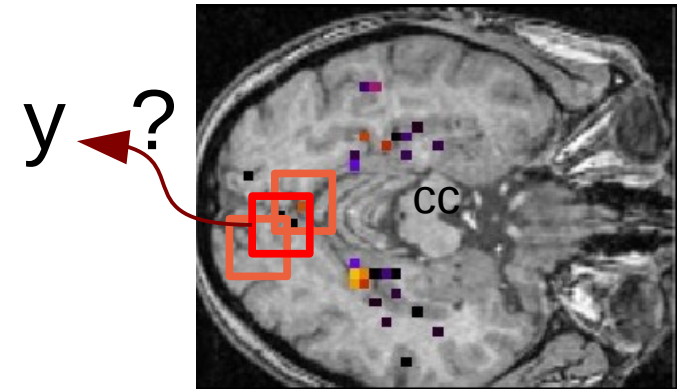


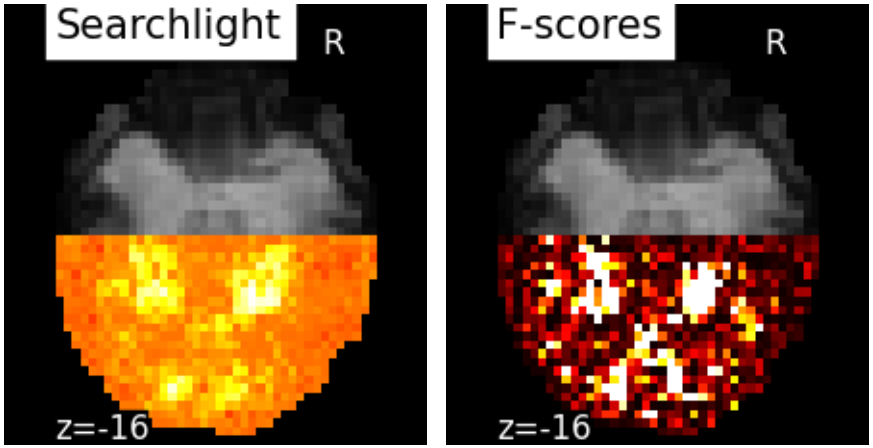
[Knops et al., 2009]

The searchlight method

[Kriegeskorte et al. PNAS 2006]

- Prediction based on a small group of neighboring voxels
- Change the region location
- Obtain a map of statistical association





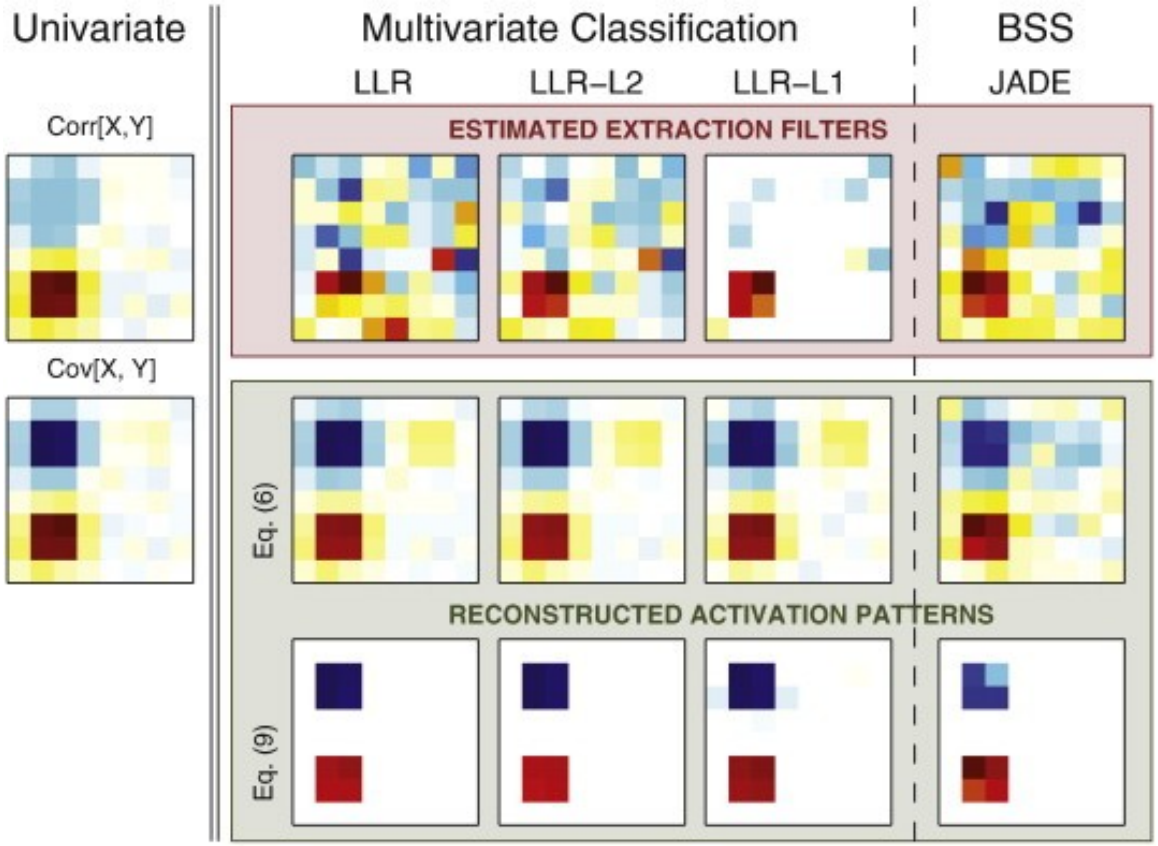
Drawbacks:
- Cost (loop over voxels + permutation testing)
- Advantage wrt standard brain mapping ?

(Non-)identifiability of the model $W = W_{true}$?

$$\hat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} \mathcal{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}) + \alpha \mathcal{R}(\mathbf{w})$$

- [Haufe et al. Neuroimage 2013] In decoding, you don't recover the true pattern but a **filter**
- The encoding model violates the conditions for good reconstruction [Varoquaux et al. 2012]
- Better support recovery by introducing relevant priors on the decoder [Varoquaux et al. 2012]
  - Sparsity
  - Block structure

Non-identifiability of the model ?



✗ MVPA **cannot** recover the true sources ;

➔ aims at finding a good discriminative model (“filters”)

➔ not at estimating the signal.

✗ A correction taking covariance structure is necessary...

✔ … but you cannot estimate the signal/noise covariance properly !

✔ However, this can be improved by choosing relevant priors

[Haufe et al. NIMG 2013]

Recovery...

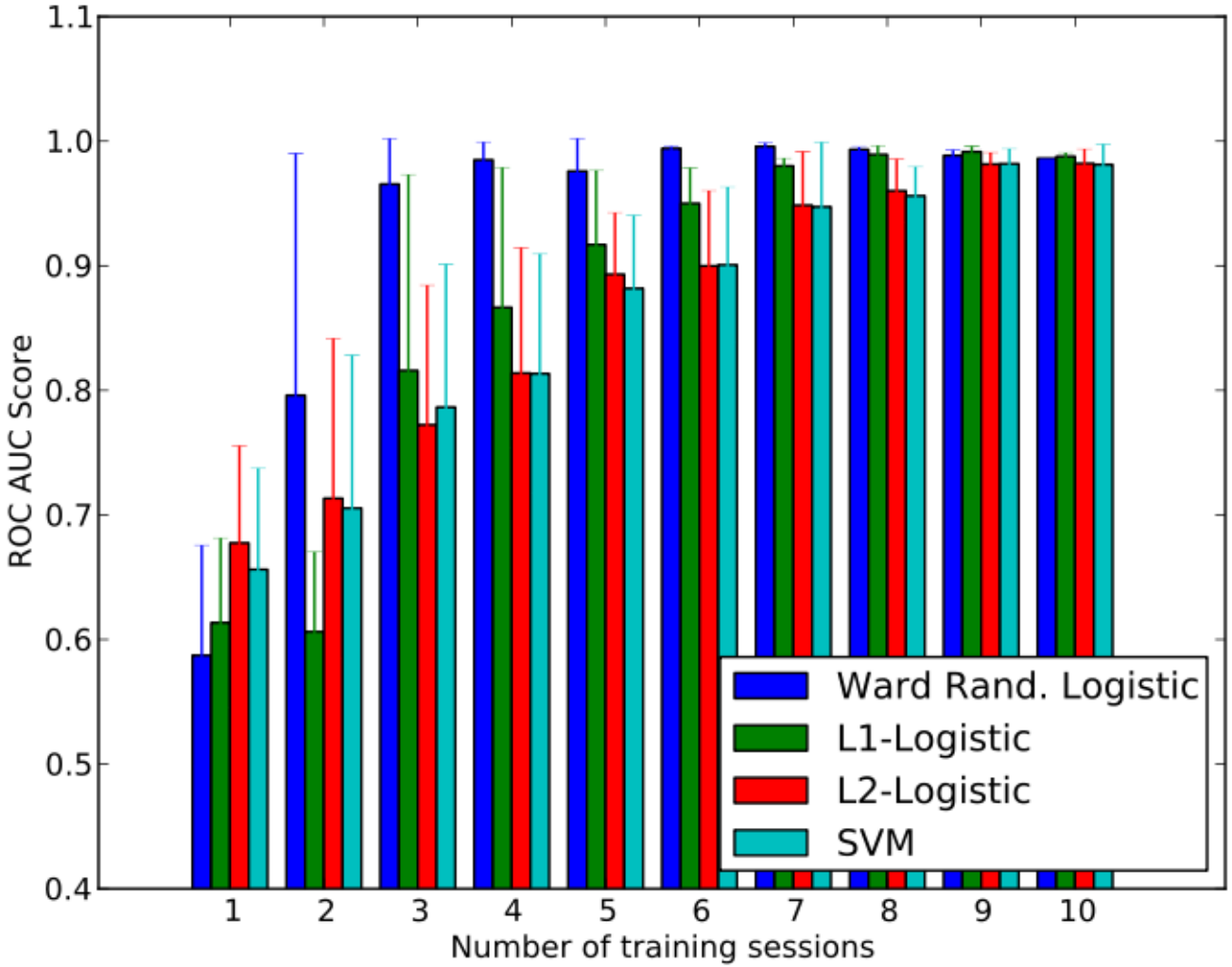
- **Compressive sensing:**
  - detection of k signals out of p (voxels)
  - with only n observations << k
  - **Problem: neuroimaging data are correlated (spatially)**
    - Violate *Restricted Isometry property.*

$$\exists \delta \in [0,1] : \forall \mathbf{w} \in \mathbb{R}^p, (1-\delta)\|\mathbf{w}\| \leq \|\mathbf{X}\mathbf{w}\| \leq (1+\delta)\|\mathbf{w}\|$$

For **any** submatrix X of the neuroimaging data matrix

How to measure the recovery of the set of regions ?
How to improve recovery?

Measuring recovery by subsampling



[Haxby Science 2001] dataset:

Trying to discriminate faces vs houses: level of performance achieved with limited number of samples

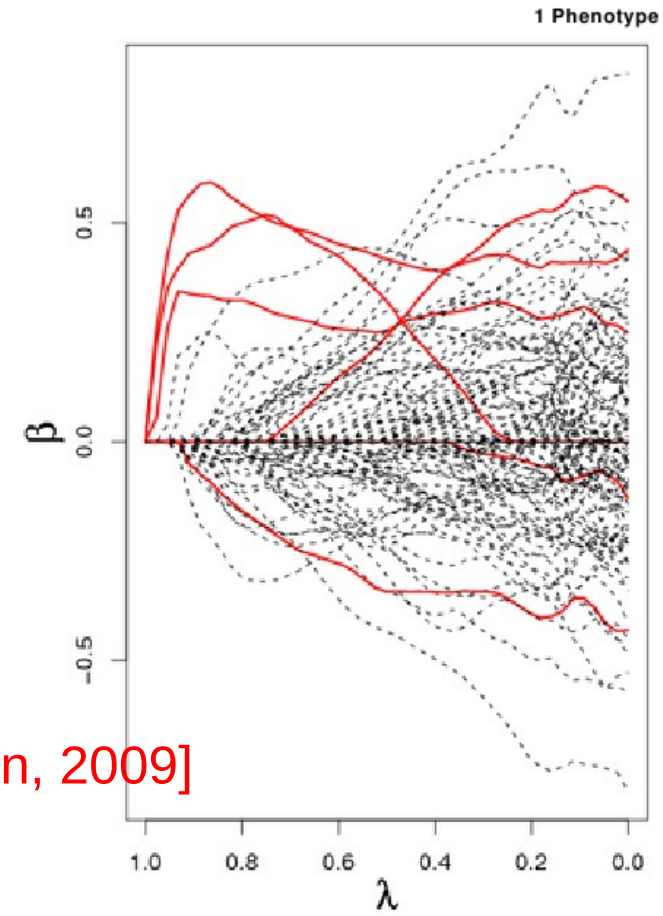
Solution: bagging to better identify

$$\hat{\mathbf{w}}^{lasso} = \text{argmin}_{\mathbf{w}\in\mathbb{R}^p} \|y - \mathbf{X}\mathbf{w}\|^2 + \alpha\|\mathbf{w}\|_1$$

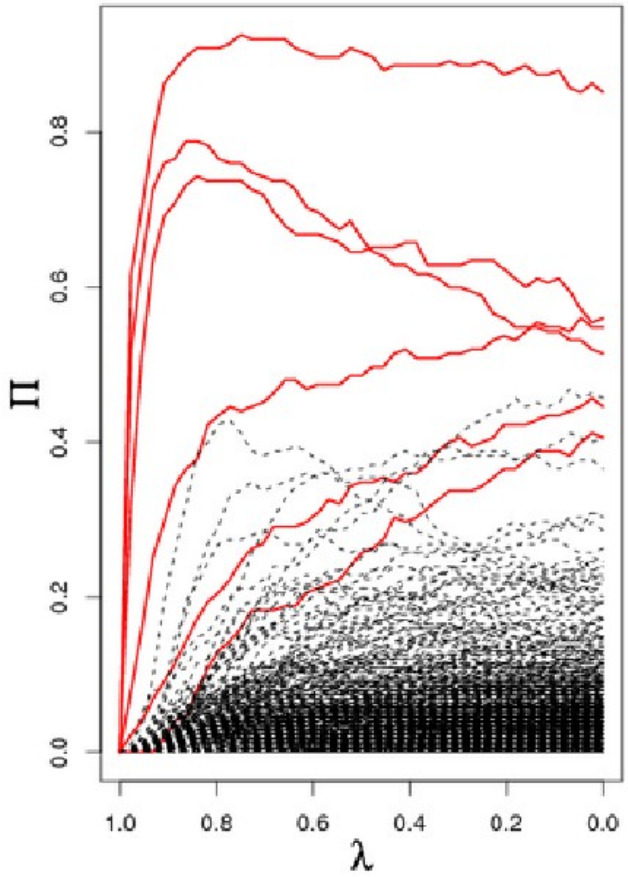
- Stability selection = randomization of the features + bootstrap on the samples
- Improved feature recovery... for **few, weakly correlated** features

[Meinshausen and Bühlman, 2009]

Lasso path



stability path of Lasso



Hierarchical clustering and randomized selection

Cluster-based stability selection algorithm

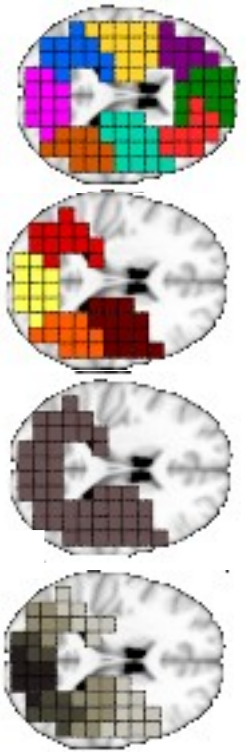
(1) **Loop**: randomly perturb the data

(2) Ward agglomeration to form q features

(3) sparse linear model on reduced features

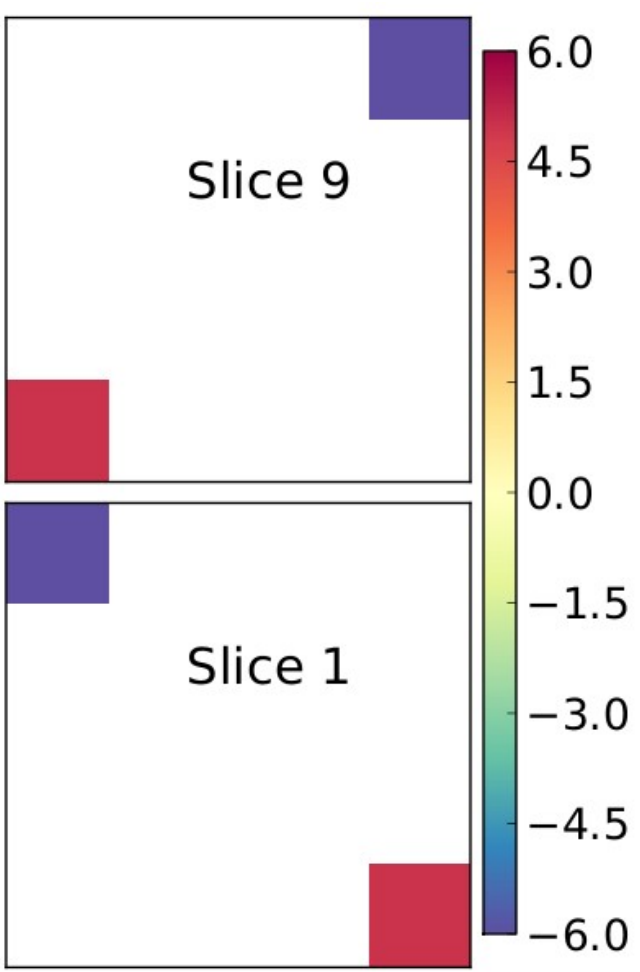
(4) accumulate non-zero features

(5) threshold map of selection counts

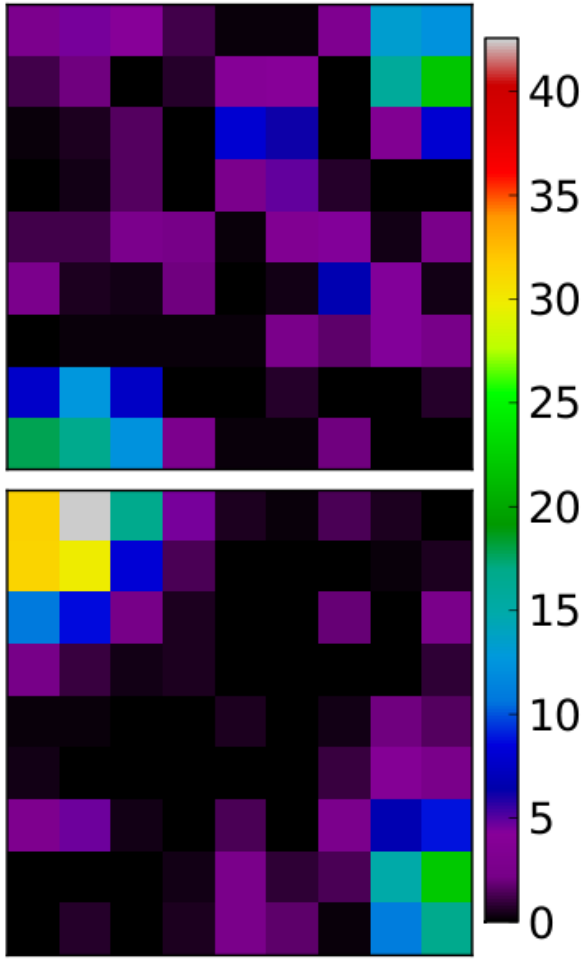


[Gramfort et al. MLINI 2011]

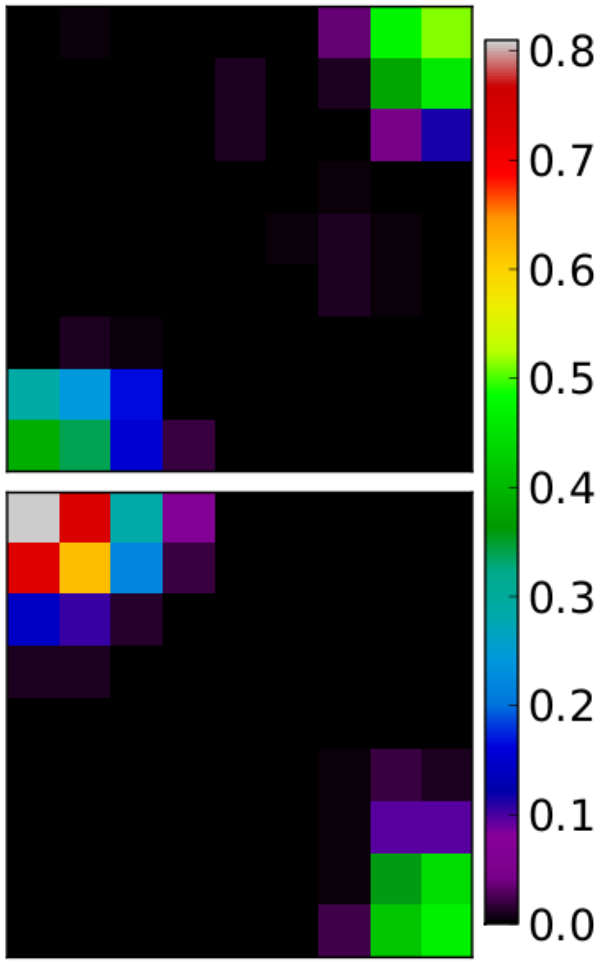
Simulation study



Ground truth



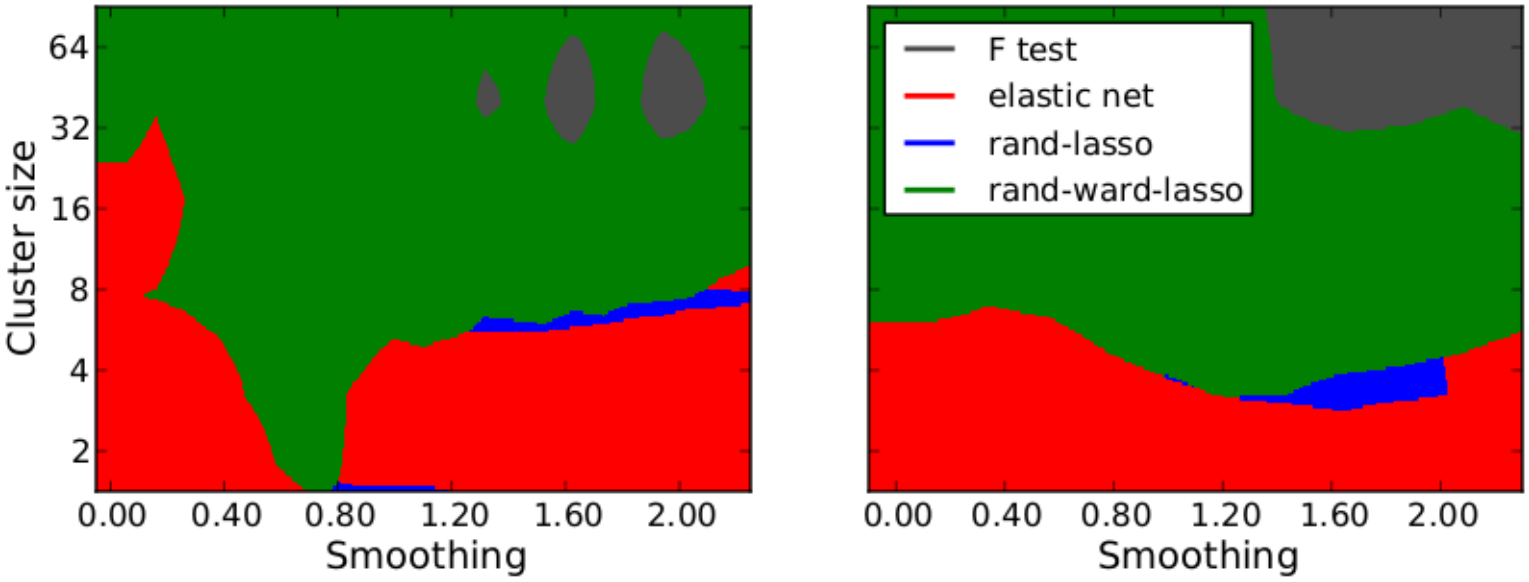
F test



Randomized Ward logistic

The best approach for feature recovery depends on the problem

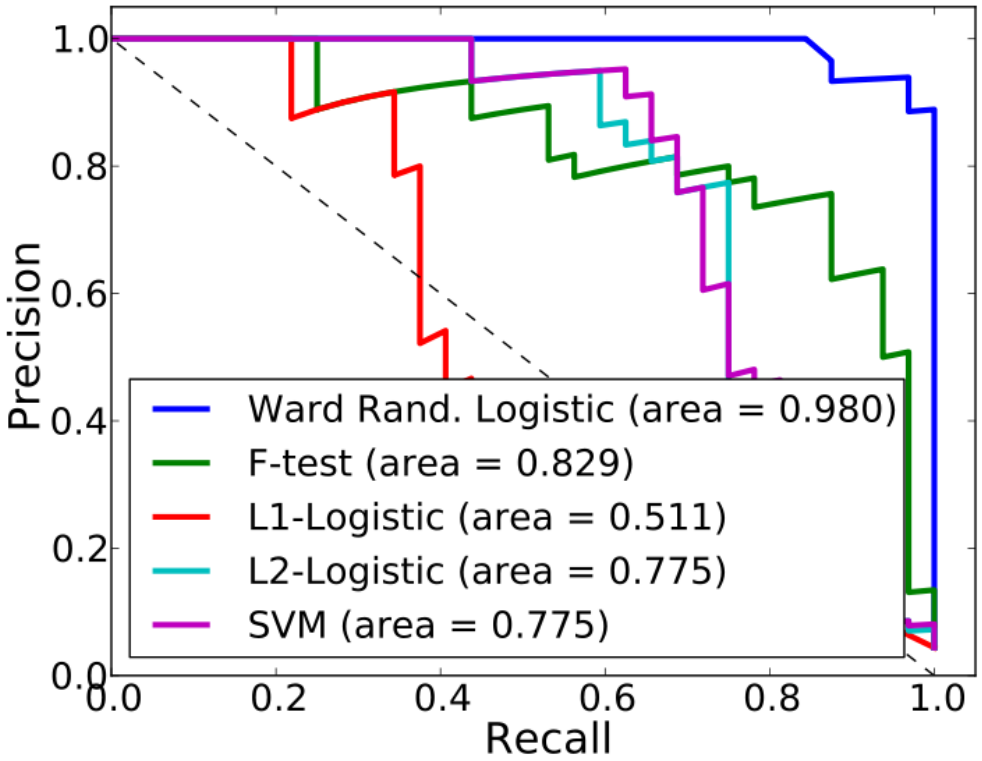
- The response depends on the characteristics of the problem: **smoothness** (coupling between signal and noise) and **clustering** (redundancy of features)



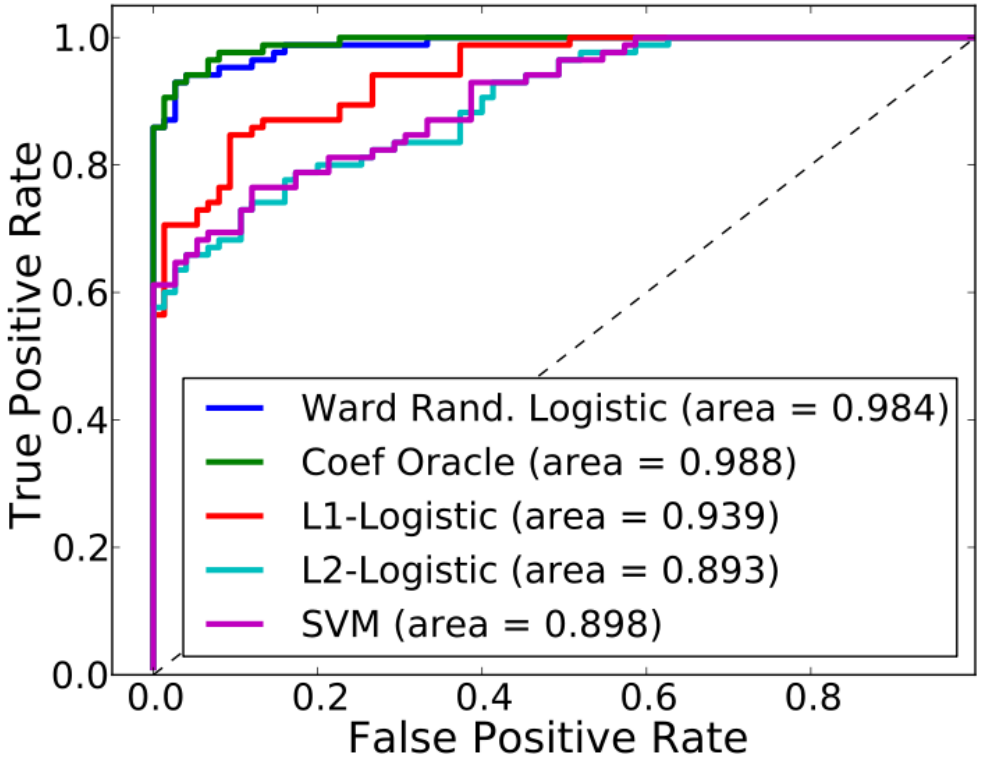
128 samples 256 samples

[Varoquaux et al. ICML 2012]

Simulation study



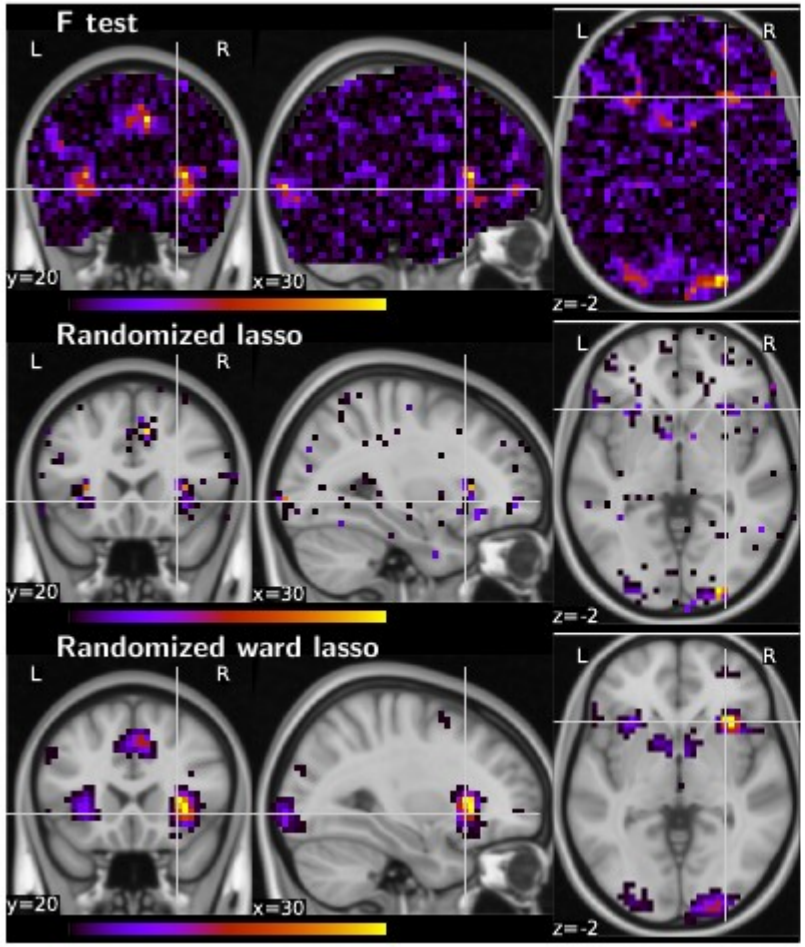
Identification accuracy



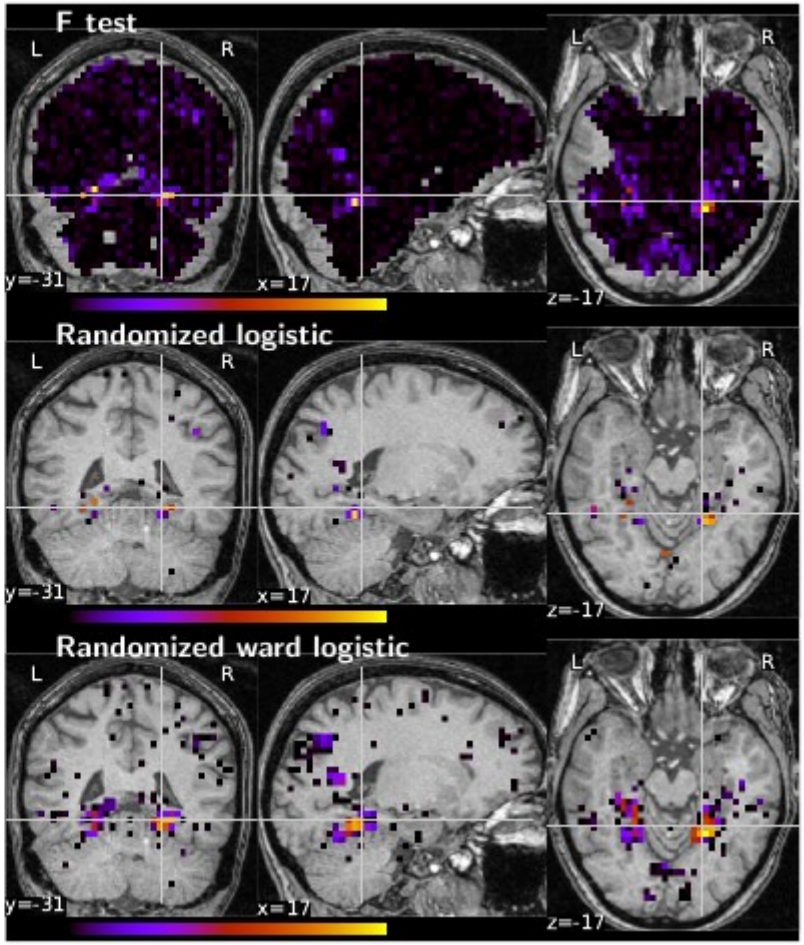
Prediction accuracy

Improves both prediction and identification !

Examples on real data



Regression task
[Jimura et al. 2011]



Classification task
[Haxby et al. 2001]

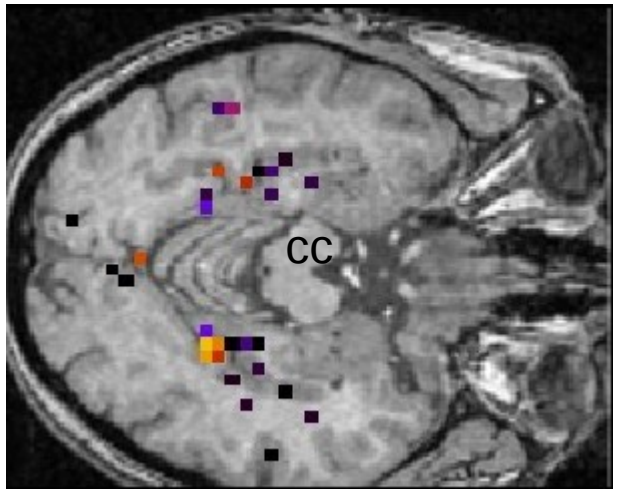
Discussion

✔ SVM and sparse models **less** powerful than univariate methods for recovery.

✔ Sparsity + clustering + randomization: good recovery

⇒ Multivariate brain mapping

✗ High computational cost (parameter setting)



Tailored priors for brain activity decoding

- Relevant priors: smoothness, sparsity, analysis sparsity
- Engineering of penalized regression models

Regularization framework

w = the discriminative pattern

Constrain w to select few parameters that explain well the data.

→ Penalized regression

$$\hat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} \mathcal{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}) + \alpha \mathcal{R}(\mathbf{w})$$

- L(y, X,w) is the *loss function*
- αR(w) is the **penalization** term, α > 0

$\alpha\mathcal{R}(\mathbf{w}) = \alpha\|\mathbf{w}\|^2$		Ridge (no sparsity)
$\alpha\mathcal{R}(\mathbf{w}) = \alpha\|\mathbf{w}\|_1$		Lasso (very sparse)
$\alpha\mathcal{R}(\mathbf{w}) = \alpha\left(\rho\|\mathbf{w}\|_1 + (1-\rho)\|\mathbf{w}\|^2\right)$		Elastic net (sparsity + grouping)
$\alpha\mathcal{R}(\mathbf{w}) = \alpha\left(\rho\|\mathbf{w}\|_1 + (1-\rho)\|\nabla\mathbf{w}\|^2\right)$		Smooth lasso (sparsity + smoothness)
$\alpha\mathcal{R}(\mathbf{w}) = \alpha\left(\rho\|\mathbf{w}\|_1 + (1-\rho)\|\nabla\mathbf{w}\|_{2,1}\right)$		Total variation (piecewise sparsity)

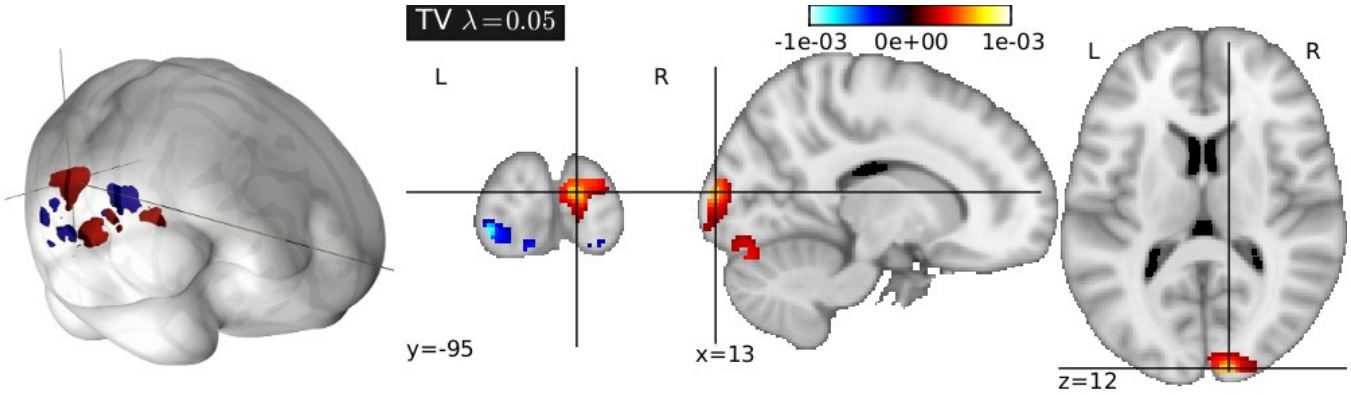
Priors and penalization: Brain decoding = engineering problem ?

Prior on the relevant activation maps

Penalization in regularized regression

Design of a norm $\|w\|$ to be minimized

Example: Total Variation penalization [Michel et al. 2011]

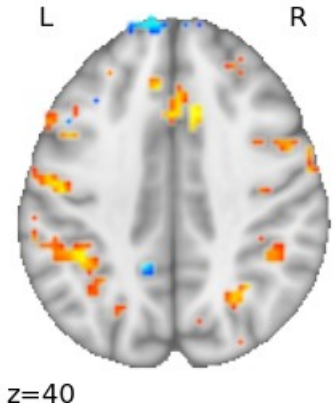


Do we need to bother about sparsity ?

Is brain activation (connectivity,..) “sparse” ? No !
But...

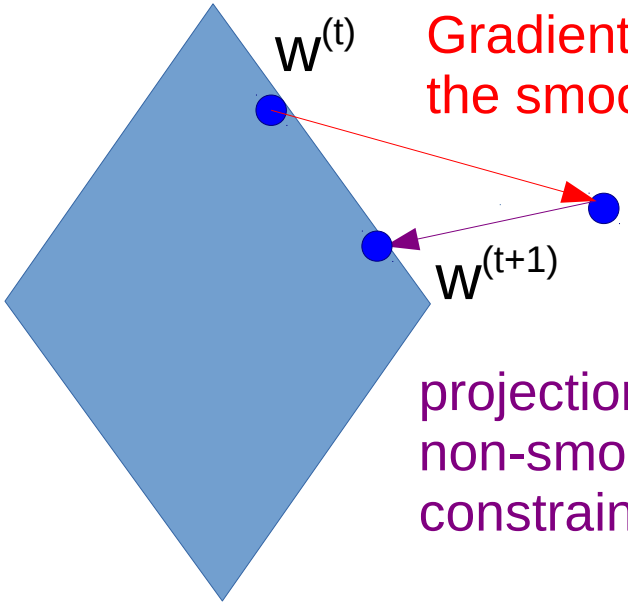
In neuroscience, people estimate discriminative patterns that look like:

But in a neuroimaging article, it will look more like



If you want to show the truly discriminative pattern, you need it to be sparse !

Solution: (F)ISTA



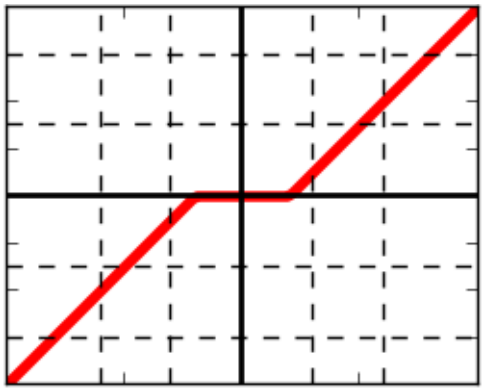
Gradient descent on the smooth terms

$$\mathbf{w}^{t+1} = \text{prox}_{\alpha\mathcal{R}}(\mathbf{w}^t - \frac{1}{L}\nabla\mathcal{L}(\mathbf{w}^t))$$

$$\text{prox}_{\alpha\mathcal{R}}(\mathbf{w}) = \text{argmin}_w \|\mathbf{w} - w\|^2 + \alpha\mathcal{R}(w)$$

projection on the non-smooth constrains

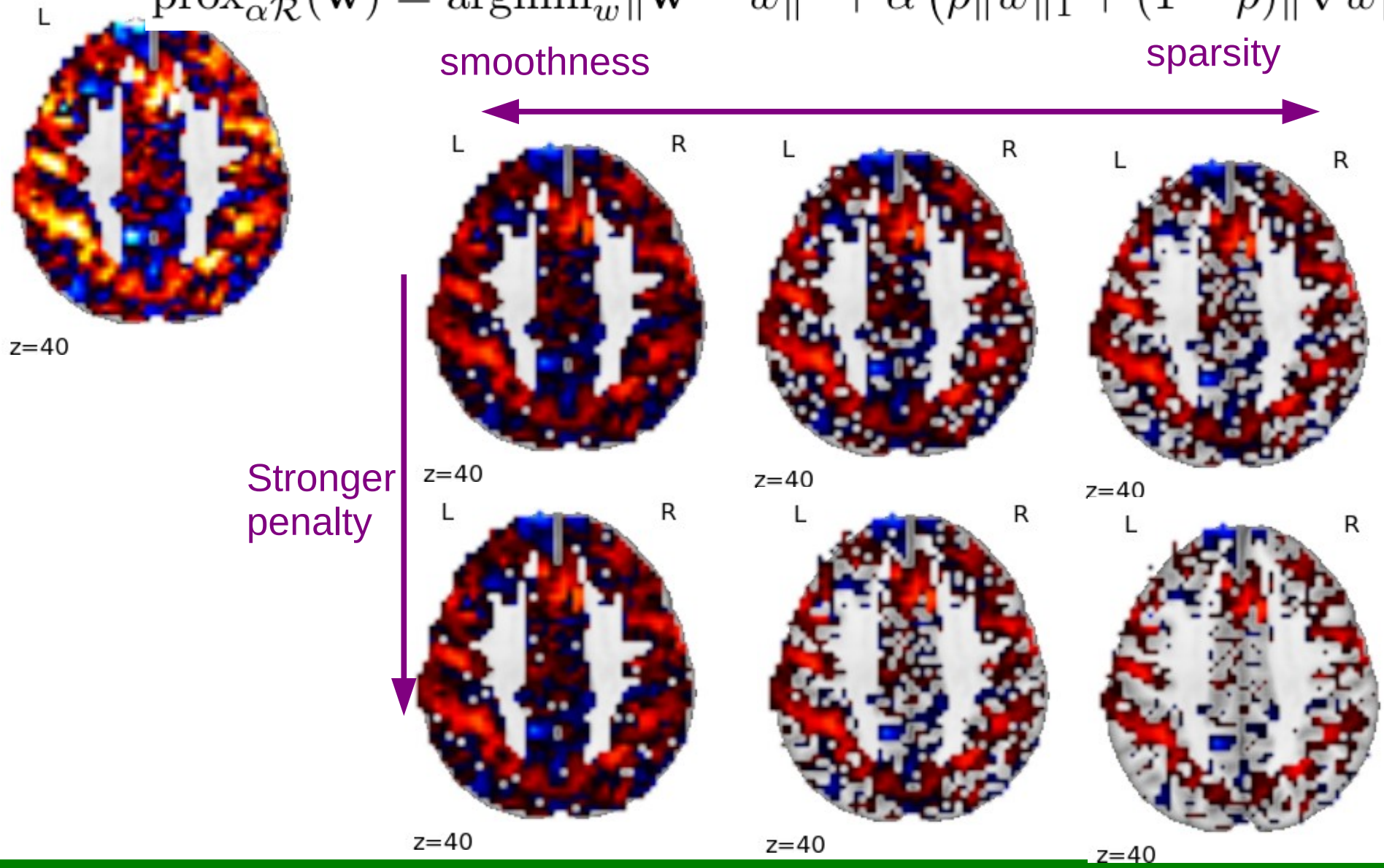
Lasso: the proximal operator is simply soft-threshodling



FISTA = accelerated ISTA (much faster convergence)

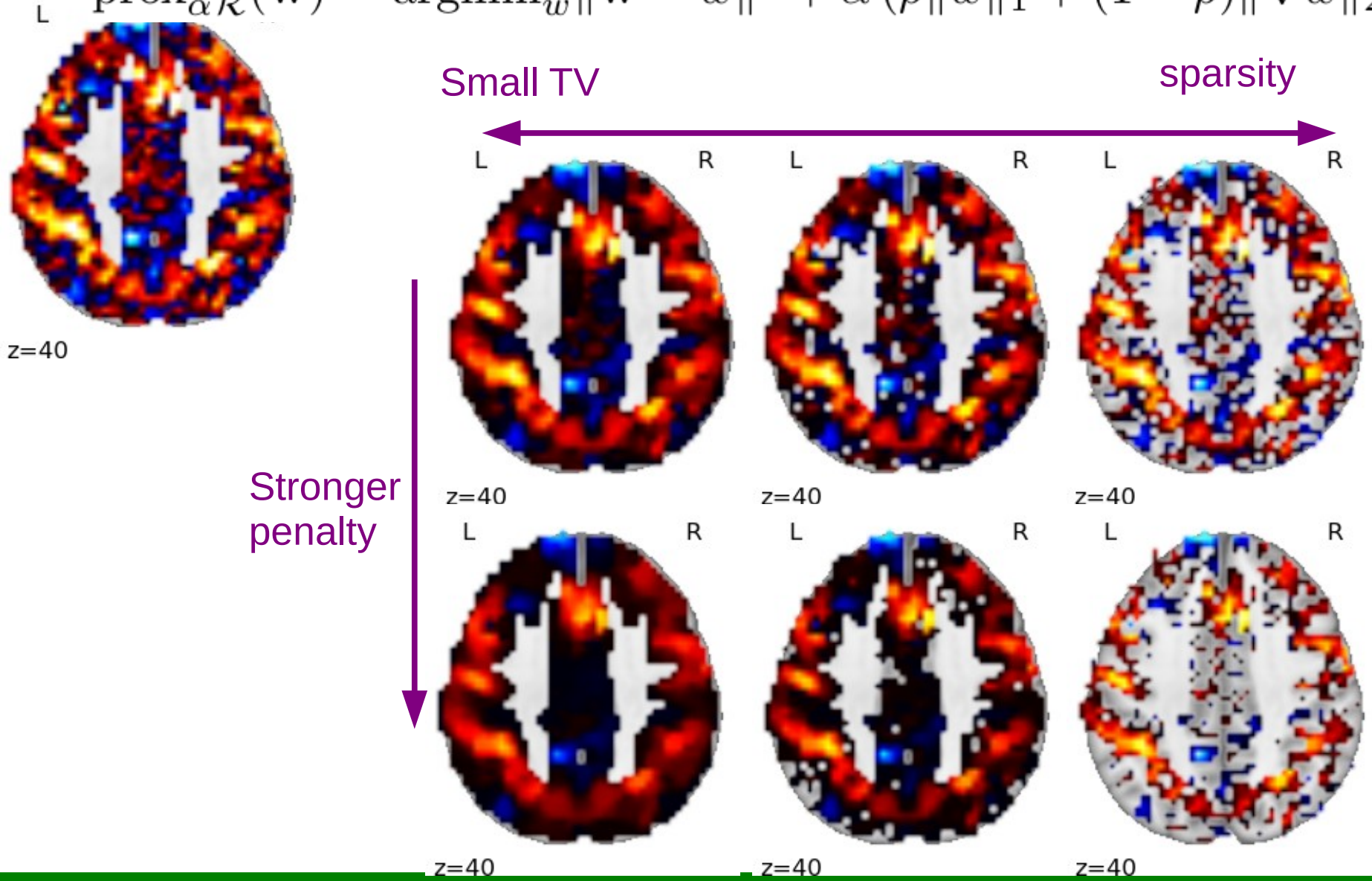
Smooth lasso: the proximal operator

$$\mathrm{prox}_{\alpha\mathcal{R}}(\mathbf{w}) = \mathrm{argmin}_w \|\mathbf{w} - w\|^2 + \alpha\left(\rho\|w\|_1 + (1-\rho)\|\nabla w\|^2\right)$$



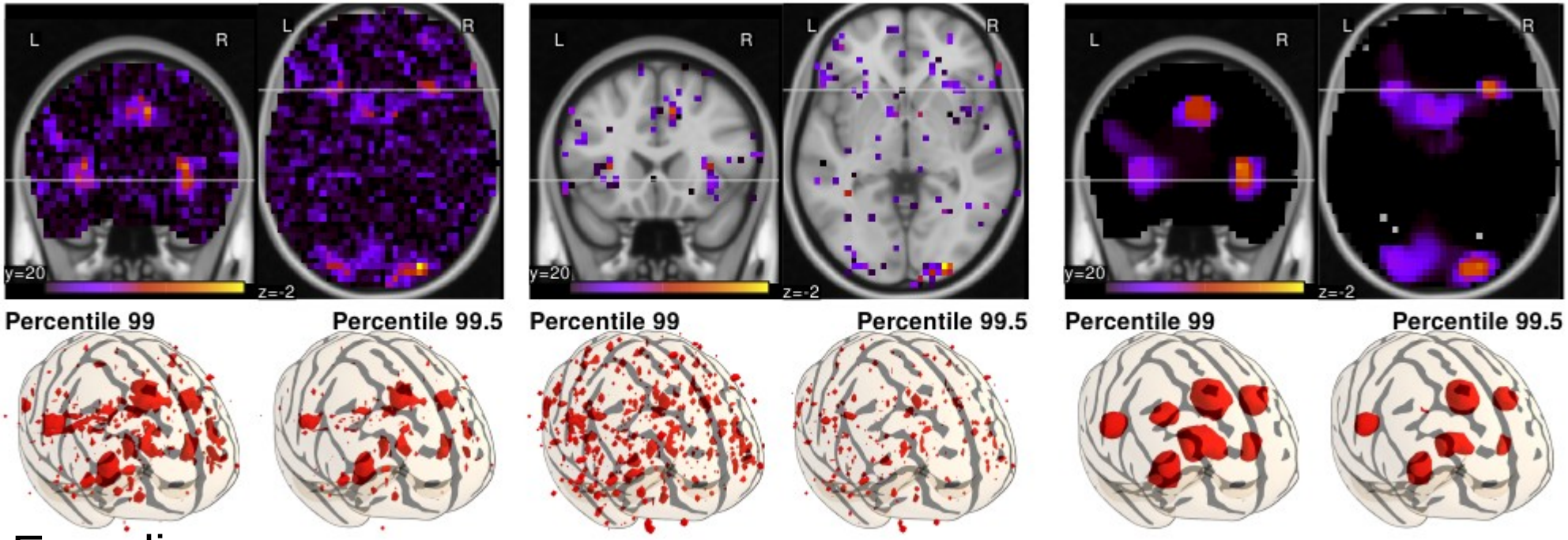
Sparse total variation: proximal operator

$$\text{prox}_{\alpha\mathcal{R}}(\mathbf{w}) = \text{argmin}_w \|\mathbf{w} - w\|^2 + \alpha\left(\rho\|w\|_1 + (1-\rho)\|\nabla w\|_{2,1}\right)$$



Example on real data

Can nevertheless be improved with adapted techniques



Encoding

Elastic net decoding

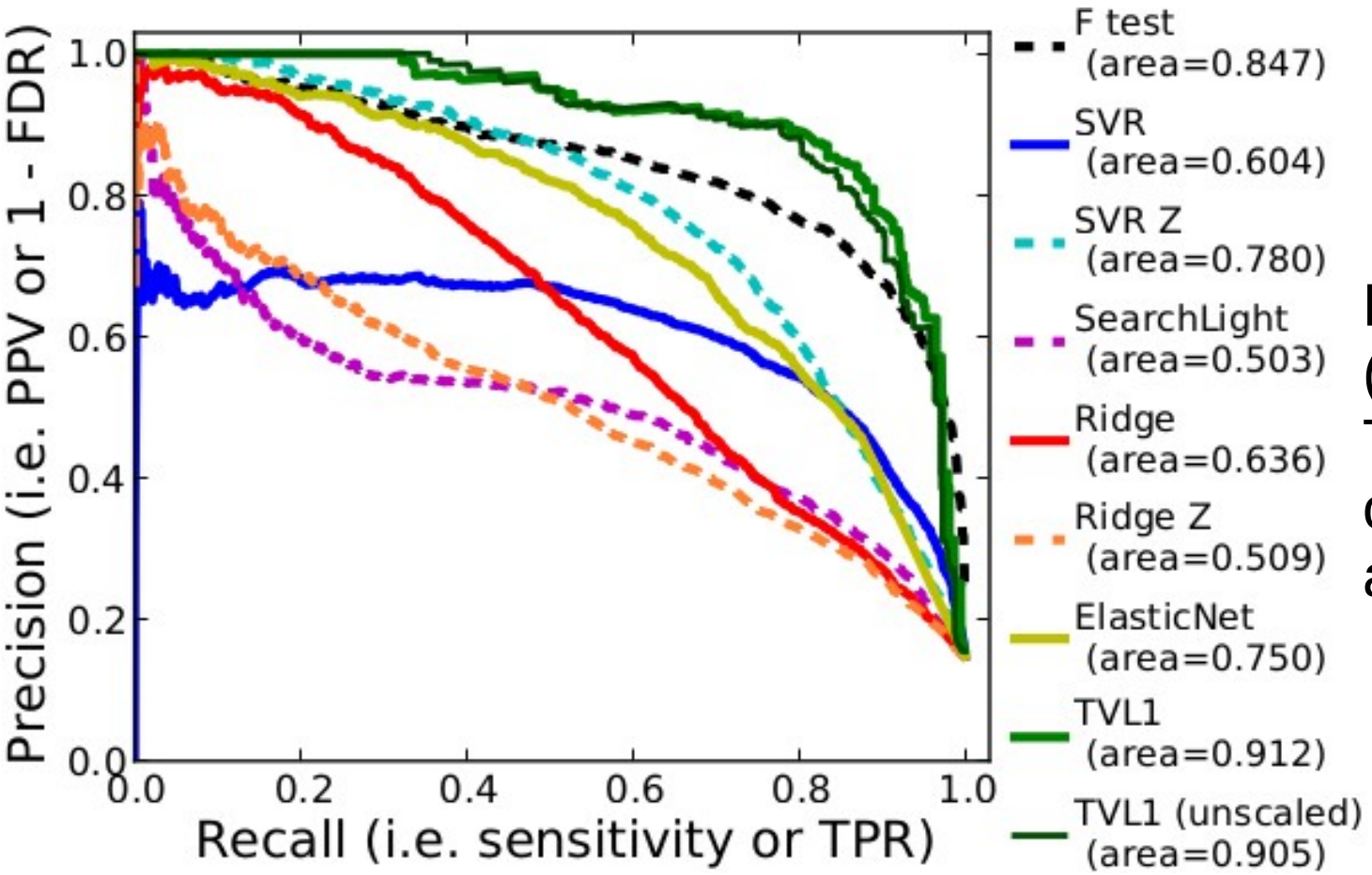
Sparse flat decoding

$$\min_{\mathbf{w}} (\mathbf{y} - \mathbf{X}\mathbf{w})^2 + \alpha \left(\rho \|\mathbf{w}\|_1 + (1-\rho)\|\mathbf{w}\|^2\right)$$

$$\min_{\mathbf{w}} (\mathbf{y} - \mathbf{X}\mathbf{w})^2 + \alpha \left(\rho \|\mathbf{w}\|_1 + (1-\rho)\|\nabla\mathbf{w}\|_{2,1}\right)$$

[Gramfort et al PRNI 2013]

Performance on recovery (simulation)



Example of recovery (simulated data): The TV-l1 prior outperforms alternatives

Discussion

- Do not try to implement these methods if not necessary. Technical issues [Dohmatob et al. PRNI 2015]
- Bayesian alternatives (ARD, smooth ARD) [Sabuncu et al.TMI 2012]
  - You lose the convexity
  - Empirical Bayes: adapts well to new data
- Cost of these methods
  - Convergence monitoring is hard
  - Smoothing + ANOVA selection + SVM is a good competitor...

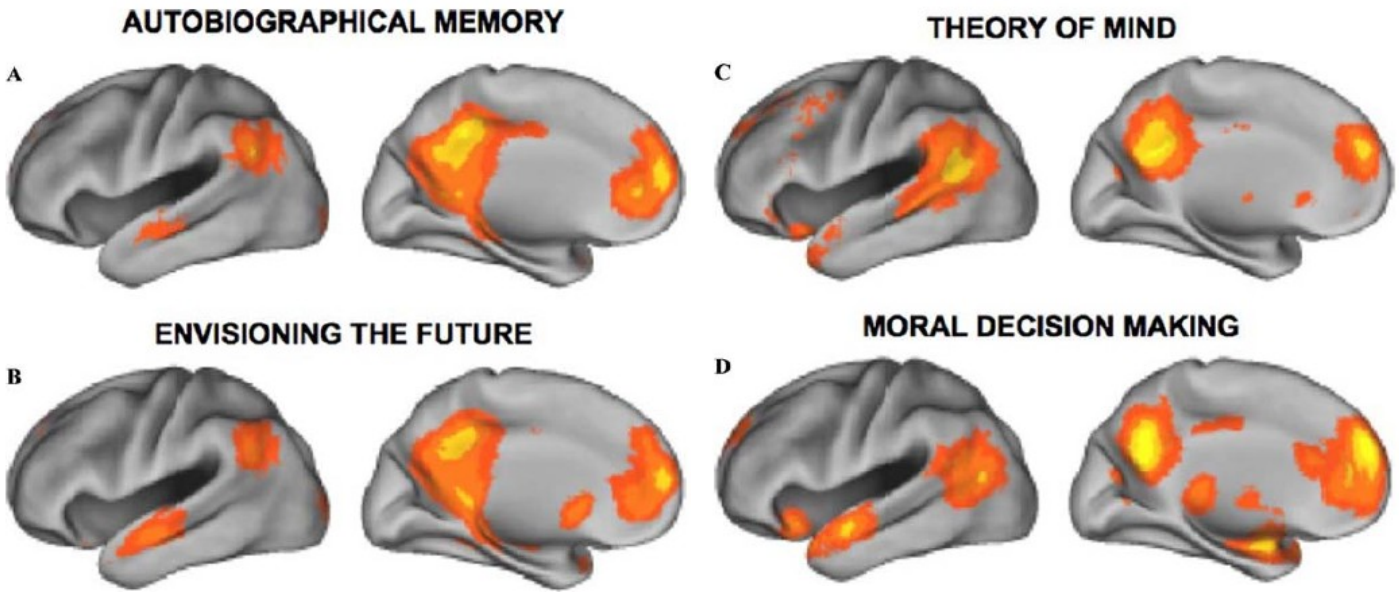
Toward Big data in cognitive neuroimaging

- Multi-label problems in cognitive neuroimaging: meta-analyses
- Mapping terms to activation maps
- Major bottlenecks

Multi-class classification is not the right setting for brain mapping

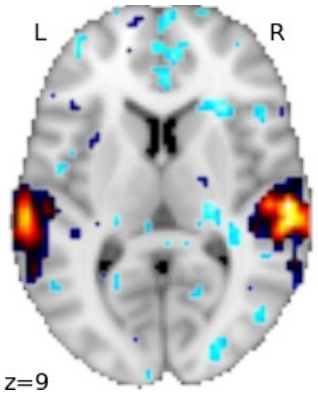
- Important questions pending....

  – Functional degeneracy between cognition and function



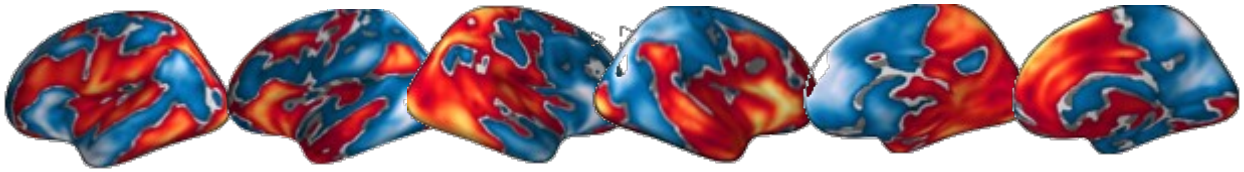
- What is this brain doing?

?



General concept

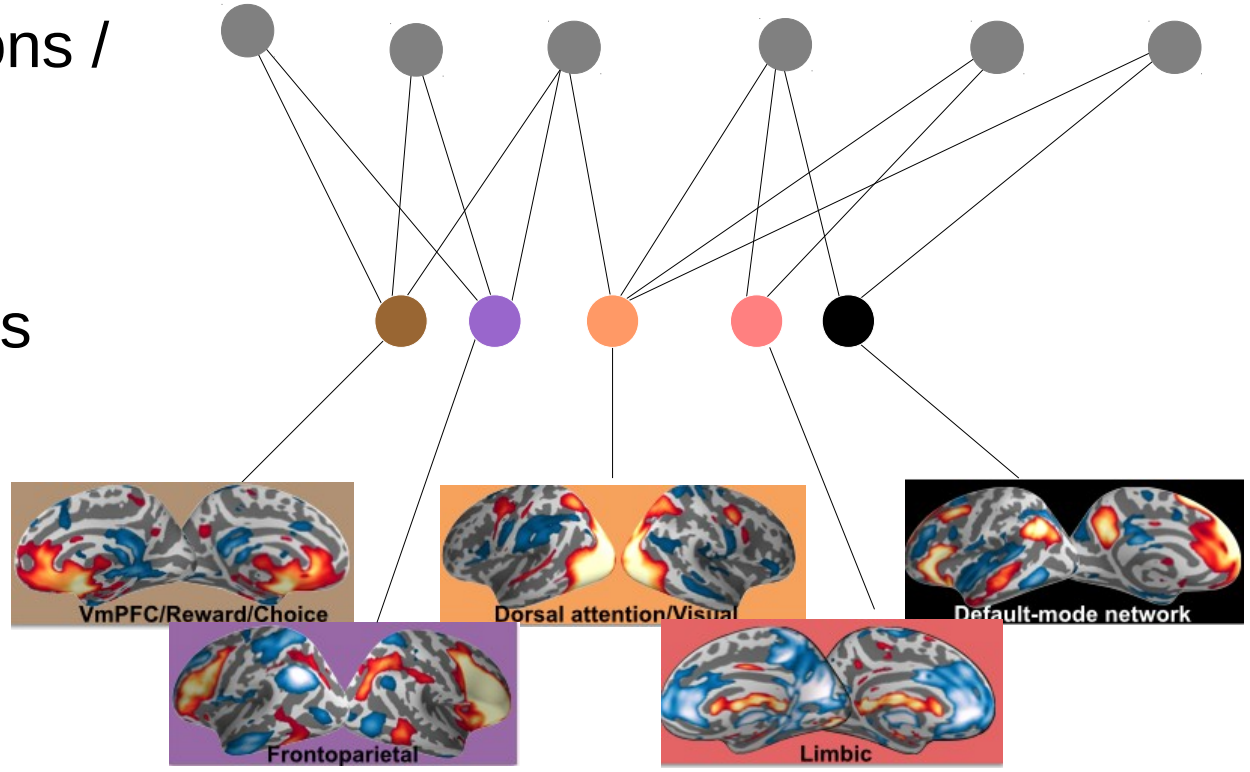
Observed SPMs



Experimental conditions /
Functional contrasts

Cognitive components

Activation patterns



Encoding

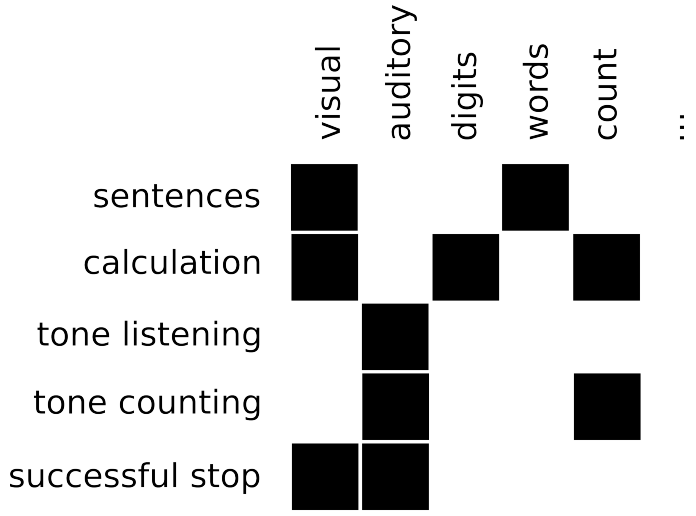
Across experiments what brain response to symbols ?

Which regions are recruited by tasks containing a given term?

- General Linear Model (GLM) for terms effects

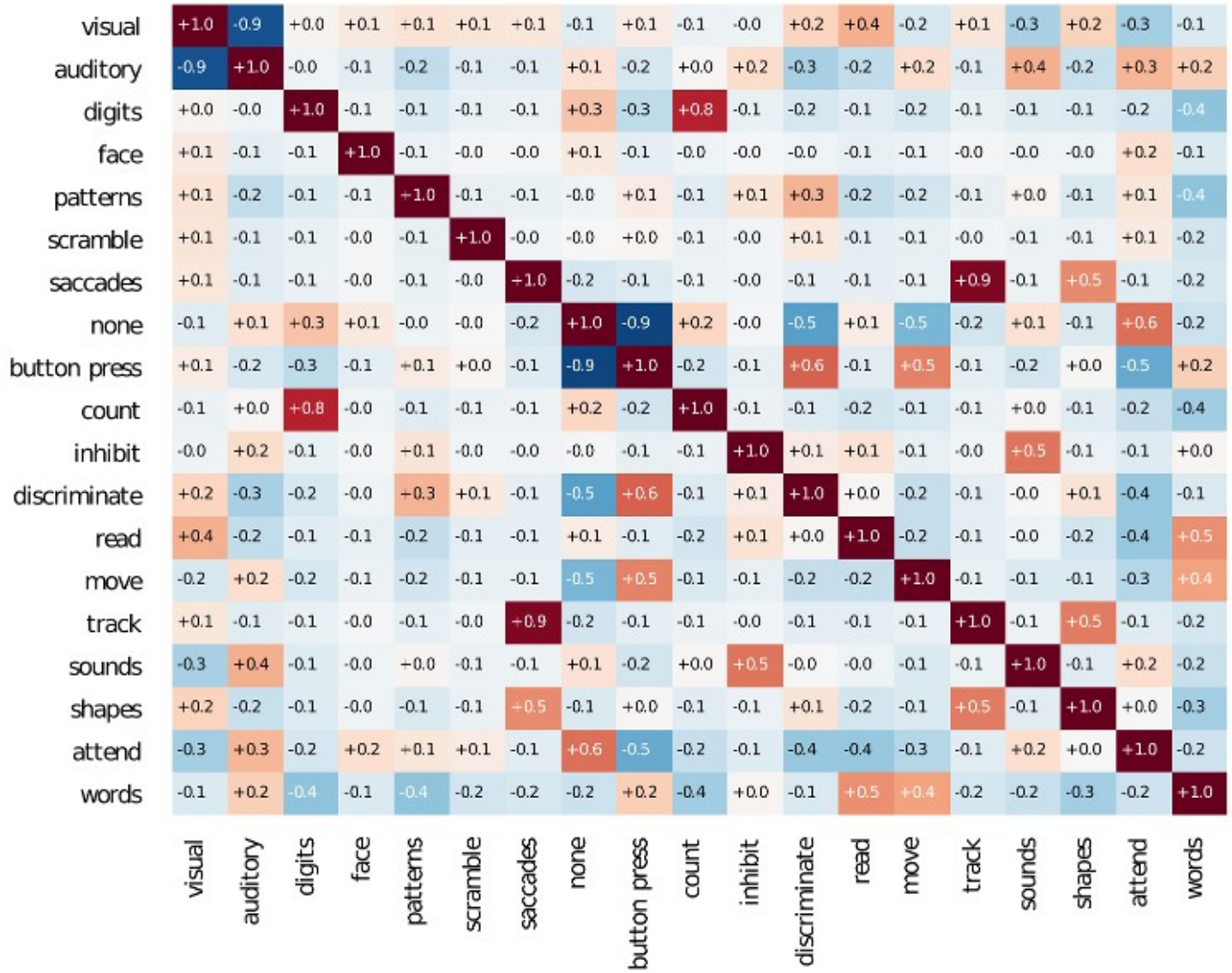
$$x = Y\beta + \varepsilon$$

x	Conditions images
Y	Design matrix
β	Terms effect
ε	Error

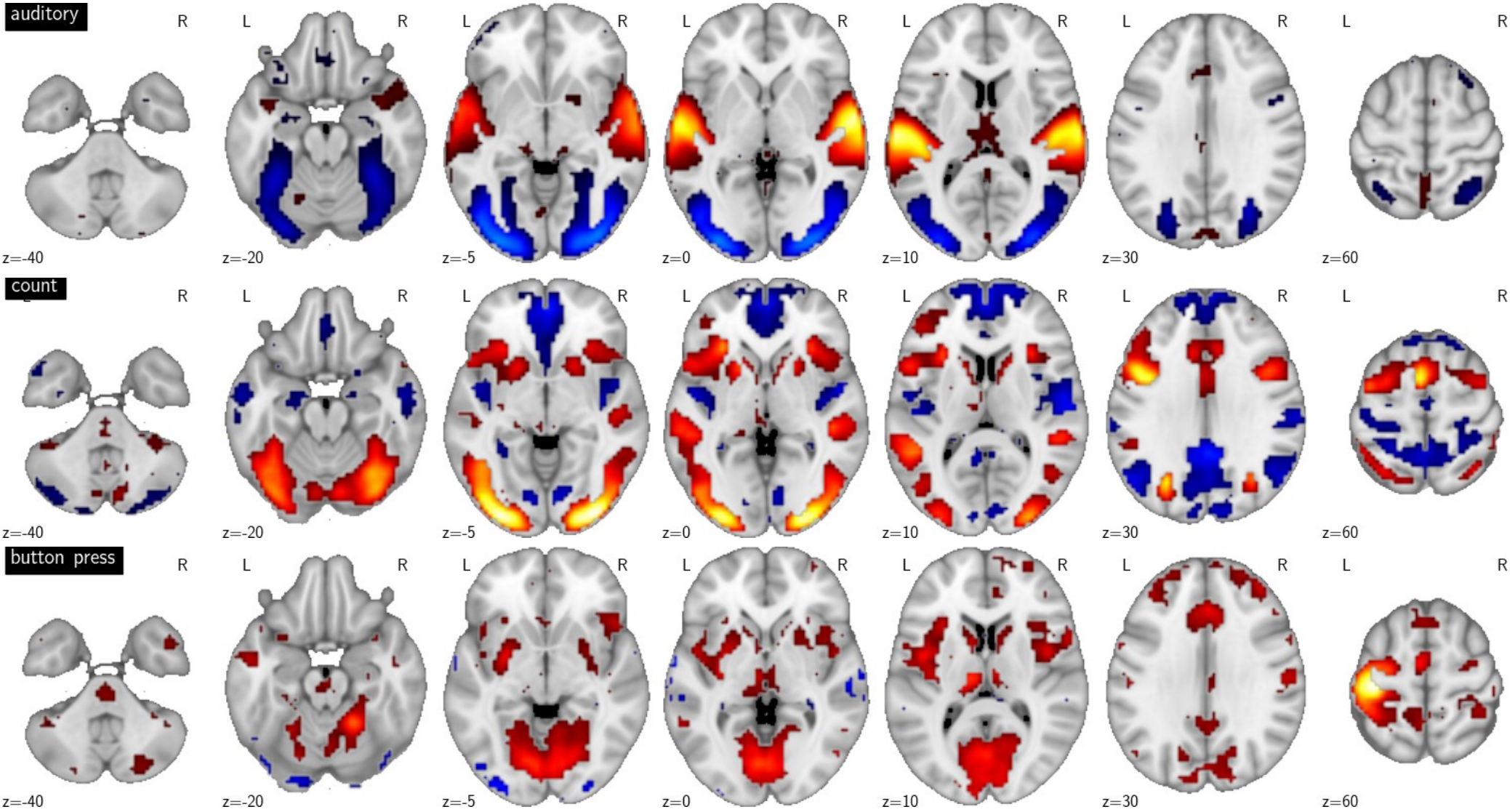


Pb with encoding approach

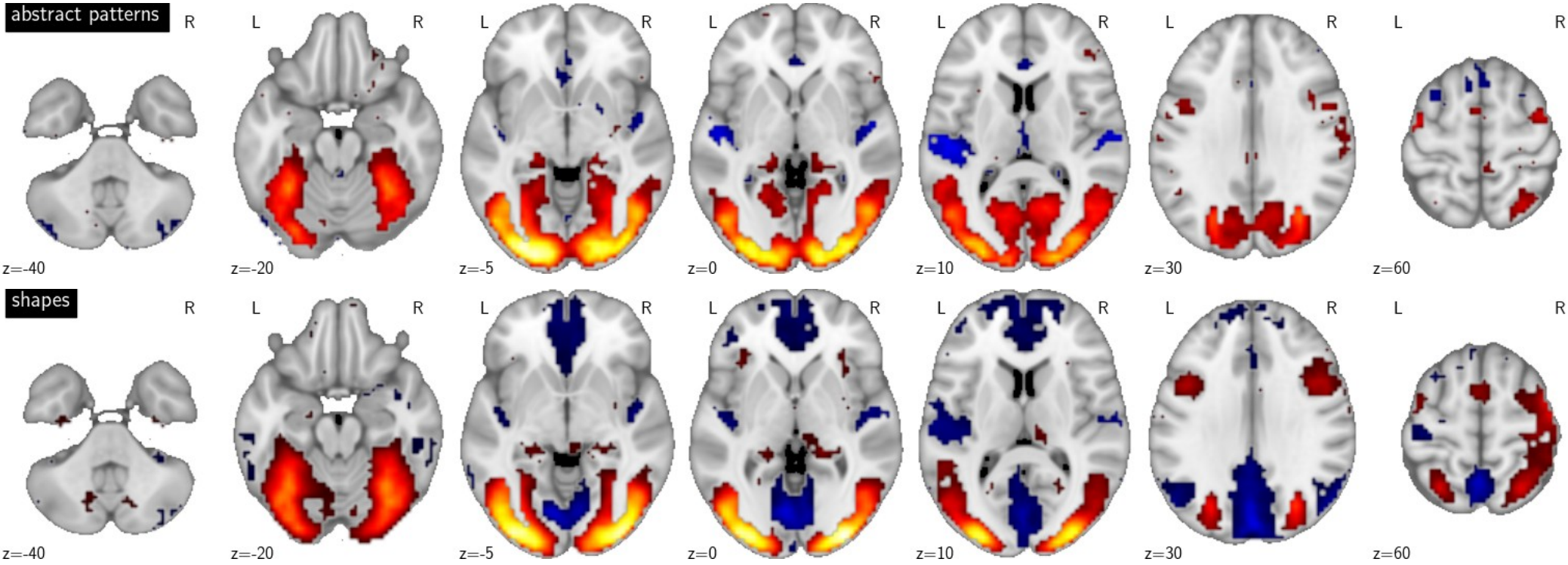
Correlation of the design matrix: difficulties from the heavily correlated terms (database bias)



Results

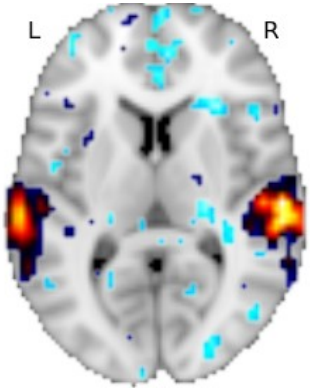


Results


abstract patterns
R
L
R
L
R
L
R
L
R
L
R
L
R
z=-40
z=-20
z=-5
z=0
z=10
z=30
z=60
shapes
R
L
R
L
R
L
R
L
R
L
R
L
R
z=-40
z=-20
z=-5
z=0
z=10
z=30
z=60

Decoding / Reverse inference

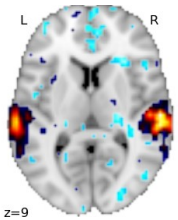
?



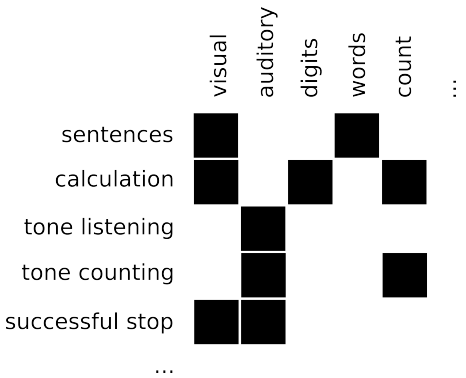
What is this brain doing?

Which regions are predictive of tasks containing a given term?

- **Multilabel** classification problem
  - more than one class may be associated with each sample
- Predict cognitive terms



Data: experimental condition images

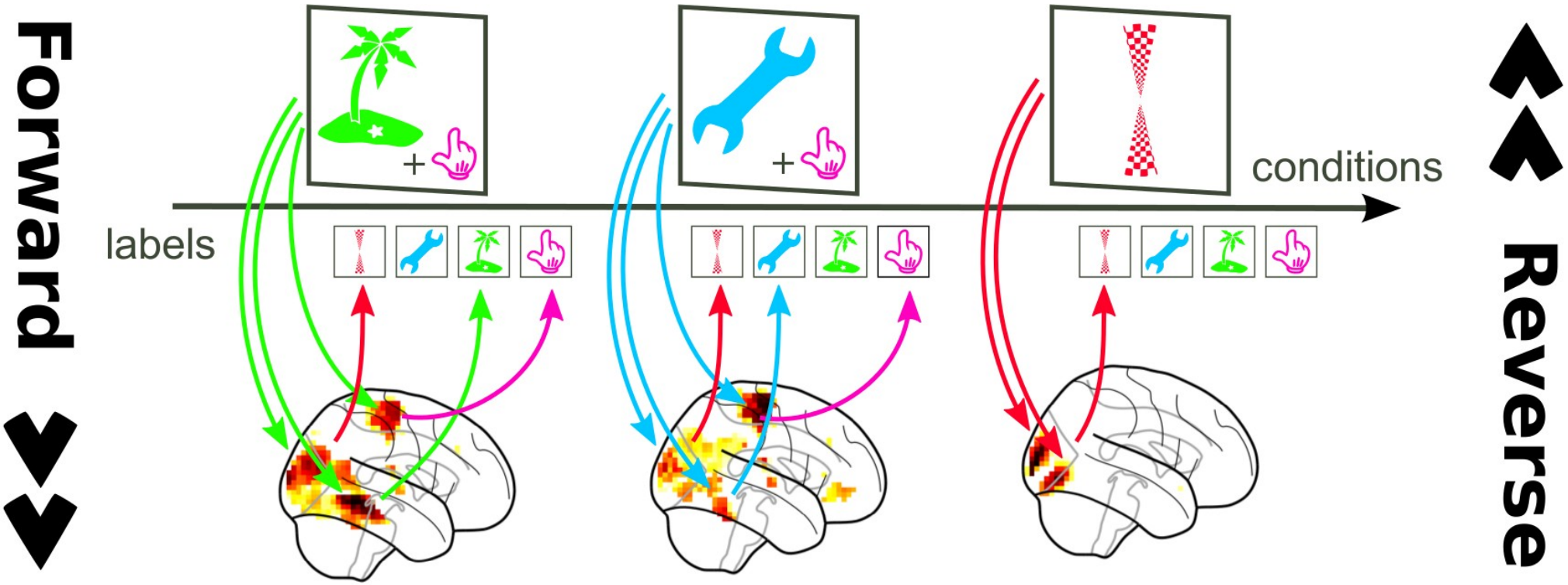


Target

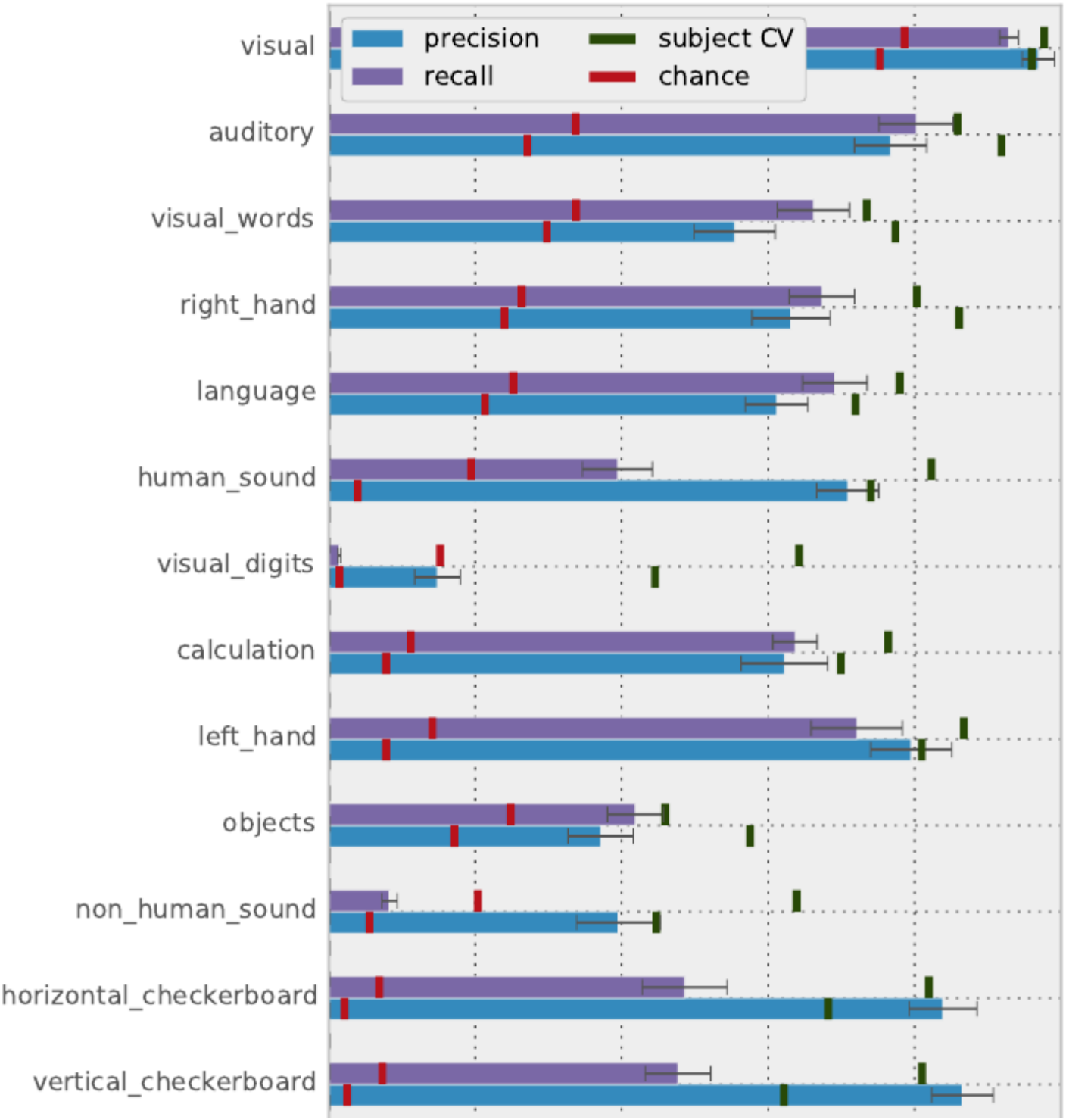
Decoding

- Problems
  - Class distribution (imbalance & covariate shift)
  - Long tailed distribution of terms
  - Standard recovery issues; noisy patterns
- Improving learning procedure
  - **Two-step** learning: discriminative sub-space selection, then prediction
  - Dimension reduction: feature clustering & Anova.
- Cross validation
  - **Leave-one study out**
  - Predict unseen conditions

Method summary



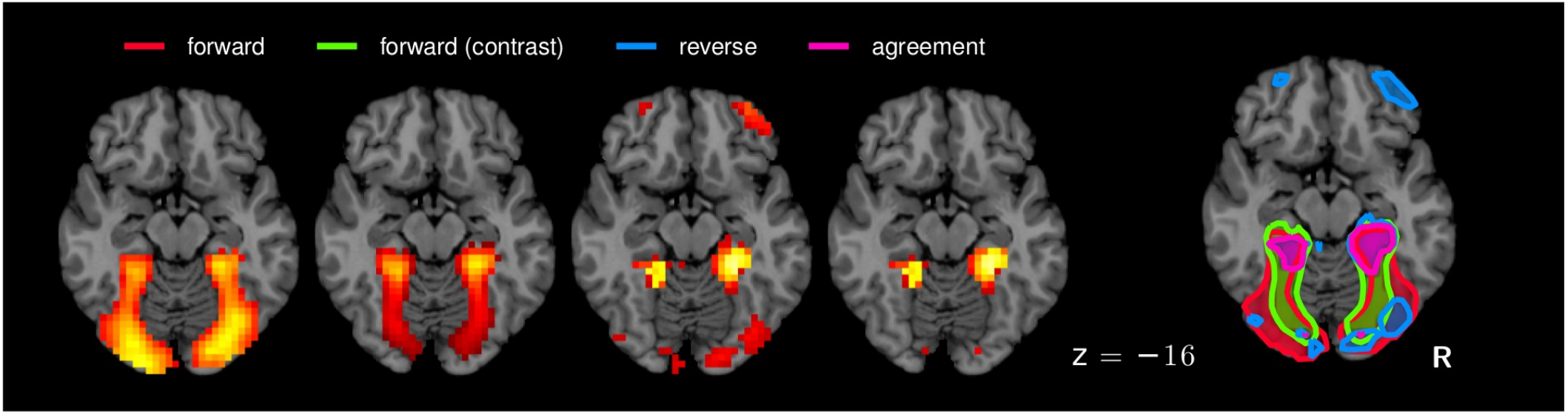
[Schwartz et al. NIPS 2013]



Precision: does not label as positive a sample that is negative

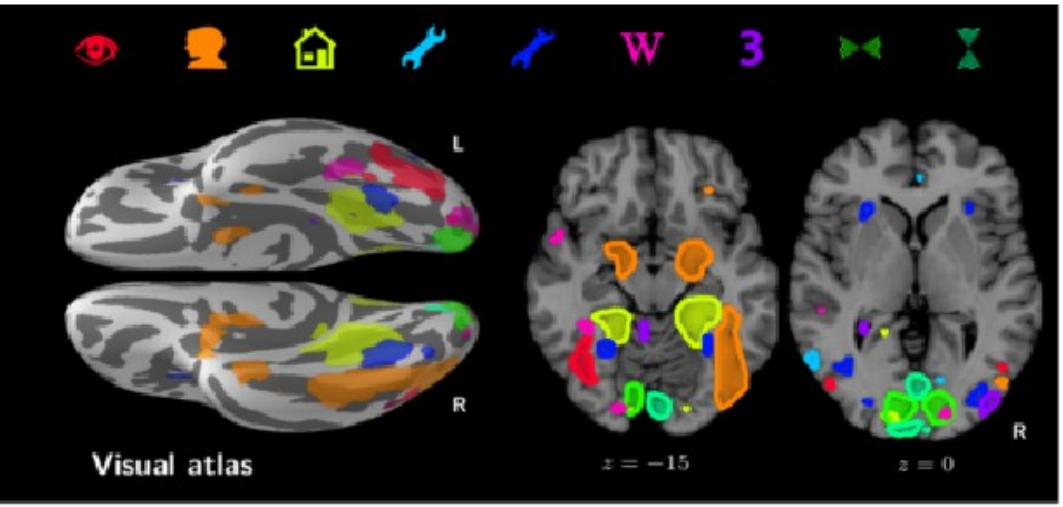
Recall: finds all the positive samples

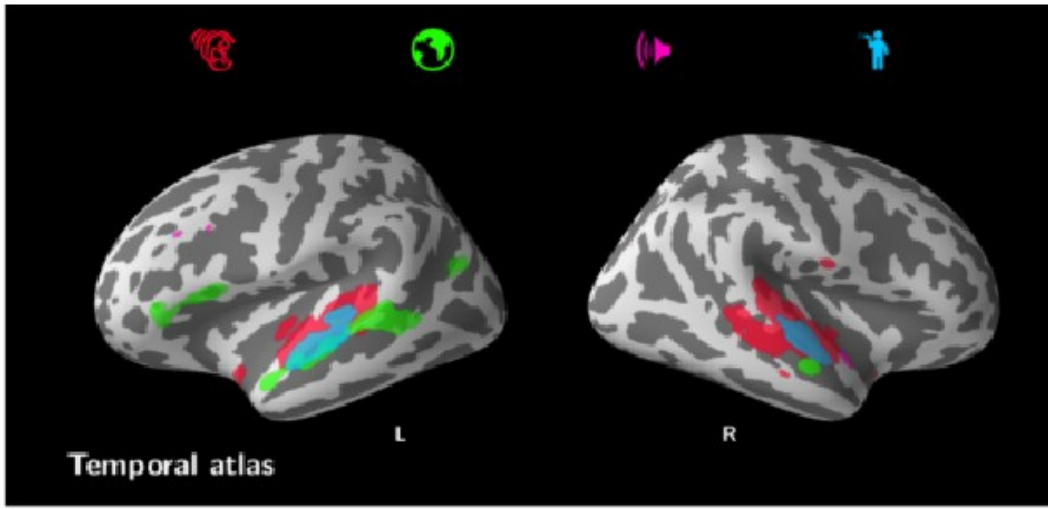
Forward and reverse inference

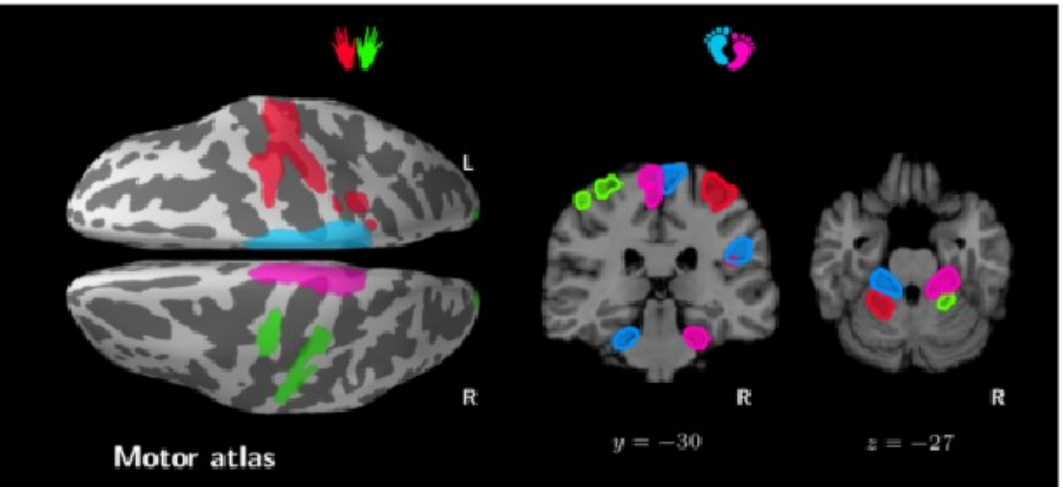


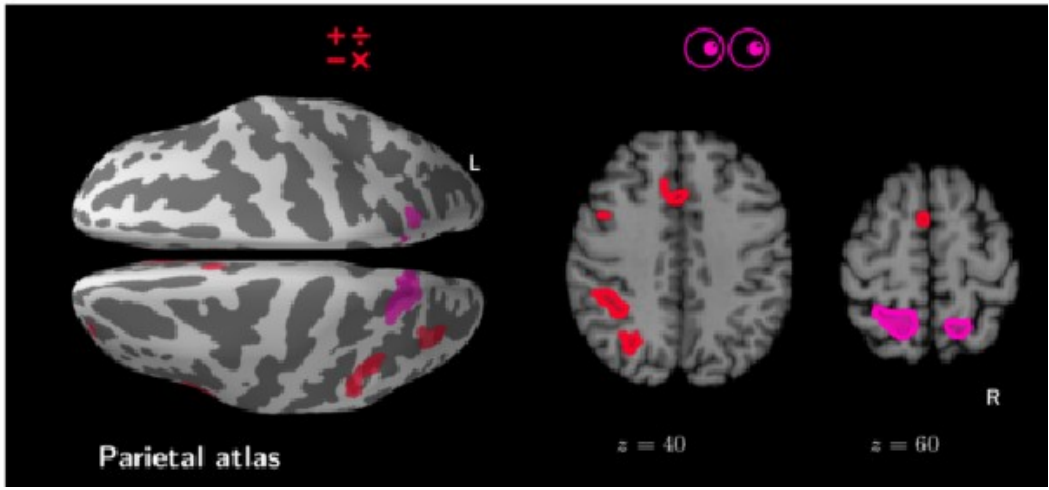
“Place”

Cognitive brain atlases









[Schwartz et al. in prep]

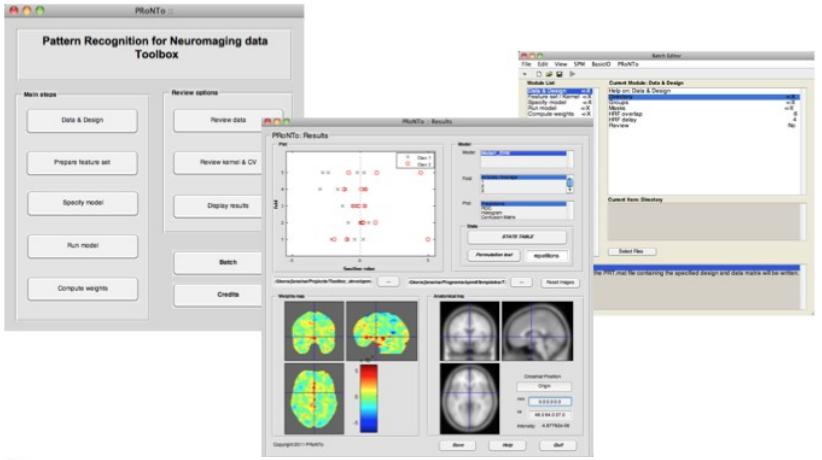
Computational bottlenecks

- Obvious limitation atm = number of training images
  - OpenfMRI and Neurovault developing fast.
- But the data handling cost. If one image = 1MB
  - 10^4 images = 10GB
  - 10^5 images = 100 GB
  - 10^6 images = 1TB
  - Moreover the resolution increases !
- Need fast learner, data compression, online training procedures. The bottleneck is memory.

[Varoquaux et al. IPMI 2013, Thirion et al. Subm.]

Available resources

Pattern Recognition for Neuroimaging Toolbox (PRoNTo)



PyMVPA: Multivariate Pattern Analysis in Python

Nilearn: Machine learning for functional neuroimaging

General Conclusions

- Look at what you do: black box models are not useful
- ML is nothing but multivariate estimation
- You probably need more data
- Do not reinvent the wheel: Pronto, PyMVPA, Nilearn
- Some topics not covered here:
  - covariance estimation (e.g. functional connectivity)
  - unsupervised models (ICA, dictionary learning...)
  - Representational similarity analysis [Kriegeskorte et al. PNAS 2008]
  - Deep nets

Thanks

Parietal
V. Michel
G. Varoquaux
A. Gramfort
F. Pedregosa,
Andés H. idrobo
J.B. Poline,
V.Fritsch,
S.Medina,
R. Bricquet
D.Bzdok
M.Eickenberg
Y.Schwartz
O. Grisel
L. Estève
M. Rahim
P.Ciuciu

Other collaborators
(thanks for the data !)
E.Eger,
R. Poldrack,
K. Jimura,
J. Haxby